Praktische statistiek voor het hoger beroepsonderwijs

J.H. Blankespoor
J.O.J. Smith

Inleiding tot de toegepaste statistiek

HB Uitgevers

Inleiding tot de toegepaste statistiek

Inleiding tot de toegepaste statistiek

vijfde, herziene druk

drs. J.H. Blankespoor
J.O.J. Smith

Praktische statistiek voor het hoger beroepsonderwijs

HB Uitgevers

HB*uitgevers*
Postbus 290
3740 AG Baarn

www.hbuitgevers.nl
e-mail: info@hbuitgevers.nl

Druk: Hentenaar Boek, Nieuwegein
Ontwerp van omslag en titelpagina: Allround Reklame & Vormgeving, Bodegraven
Tekstredactie: Karin ten Kate
Tekenwerk figuren 4.2, 6.1, 6.7, 6.8, 6.10 t/m 6.13, 9.4, 9.5 en 11.5: Blomsma Tekenbureau. Zoetermeer
Oorspronkelijke boekstijl voor Scientific Workplace: T. Hoekwater
Aangeleverd in Scientific Workplace door de auteurs
Eindopmaak: Van Gent Producties, Leiden

Digitaal gezet in Times en Arial

Vijfde druk, tweede oplage

ISBN 90 5574 239 2

Woord vooraf

'Inleiding tot de toegepaste statistiek' is een boek dat een overzicht geeft van alle basisbeginselen van de statistiek (zowel de beschrijvende statistiek als de toegepaste statistiek) met vele toepassingen. Gekozen is voor een praktische aanpak, de theorie wordt wel kort uitgelegd, maar theoretische bewijzen komen vrijwel niet voor. Het boek richt zich op het gehele hoger beroepsonderwijs, in het bijzonder de sector techniek.
Het boek is een bewerking van 'Inleiding tot de statistische analyse' van H.P. Anderson en J.H. Blankespoor, dat vier drukken beleefde en tot voor kort bij HB-uitgevers verscheen. Ook is een aantal inhoudelijke vernieuwingen doorgevoerd. Het hoofdstuk over toetsen is uitgebreid. Er is een hoofdstuk over Statistische Procescontrole toegevoegd. Er is een uitgebreide bijlage toegevoegd waarin wordt aangegeven hoe Microsoft EXCEL gebruikt kan worden om statistische problemen, zoals die in de verschillende hoofdstukken voorkomen, op te lossen.
Dit boek leent zich uitstekend voor gebruik in verschillende leeromgevingen, niet alleen in onderwijsprocessen waarin de student zelfstandig of in projecten werkt, maar ook in college-instructievorm.

juni 2001
ing. J.O.J. Smith   Brummen
drs. J.H. Blankespoor   Zoetermeer

Inhoud

1 Inleiding

1.1 Wat is statistiek?

Iedereen doet in het dagelijks leven regelmatig uitspraken op grond van zelf verrichtte waarnemingen. Iedereen spreekt wel eens toekomstverwachtingen uit op grond van gebeurtenissen die zich in het verleden hebben afgespeeld. Iedereen wordt wel eens benaderd met het verzoek in het kader van een of ander onderzoek mee te doen aan een enquête. Vrijwel iedere automobilist wordt wel eens door de politie aangehouden voor een routinecontrole van auto, autopapieren of alcoholgebruik. Iedereen wordt via de media regelmatig geconfronteerd met allerlei gegevens met betrekking tot maatschappelijk of politiek relevante zaken. Iedereen koopt voedingsmiddelen en andere levensbehoeften waarvan de kwaliteit door de fabrikant en door de keuringsdienst van waren is gecontroleerd. Velen wagen zich wel eens aan een gokje in loterij, lotto, toto, casino of welk ander kansspel dan ook. Menigeen begeeft zich op de aandelenmarkt, soms zelfs op grote schaal.
In al deze gevallen hebben we het over zaken die direct of indirect te maken hebben met het vakgebied waarover wij in dit boek het een en ander willen vertellen: statistiek.

Definitie
Statistiek is het verzamelen, ordenen, presenteren en karakteriseren van (meestal numerieke, dat wil zeggen uit getallen bestaande) informatie met als doel deze informatie te helpen analyseren en het voorbereiden van beslissingen te ondersteunen.

1.2 Wanneer gebruiken we statistiek?

In het bedrijfsleven en bij de overheid speelt statistiek een belangrijke rol. Statistiek wordt toegepast bij alle grote en vele kleinere handels- en industriële ondernemingen, vooral op het gebied van de marketing en de kwaliteitsbeheersing. In het bank- en verzekeringswezen wordt statistiek toegepast om voorspellingen te kunnen doen op korte, mid-

dellange of lange termijn. Bij vele zo niet alle gemeentelijke, provinciale en landelijke (semi)overheidsinstellingen, wordt statistiek gebruikt om informatie op logische en duidelijke wijze te verwerken en te presenteren. Aan universiteiten en hogescholen wordt statistiek vooral gebruikt om meetresultaten te interpreteren en te analyseren. In allerlei takken van de wetenschap is statistiek een waardevol hulpmiddel om bepaalde veronderstellingen te kunnen bevestigen of te ontkennen.
Bij het trekken van conclusies en het nemen van beslissingen is het niet verantwoord zich uitsluitend op intuïtie of op subjectieve inzichten te baseren. Een meer kritische houding is noodzakelijk. Statistiek is dan een waardevol hulpmiddel voor het trekken van verantwoorde conclusies. Men zal zich daarbij moeten baseren op objectieve gegevens die door middel van vooronderzoek zijn verkregen. Een dergelijk vooronderzoek kan bestaan uit een marktonderzoek, een enquête, een experiment (proefopzet), een simulatie-onderzoek (nabootsing van de werkelijkheid, meestal met behulp van een computer), een kwaliteitsonderzoek, een arbeidsanalyse, een literatuuronderzoek, enzovoorts. Zo'n onderzoek kost soms veel tijd (dus geld). De omvang van zo'n vooronderzoek is daarom meestal een afweging tussen de kosten en de te bereiken doelen. Zo is het aantal ondervraagden bij enquêtes voor het voorspellen van verkiezingsuitkomsten meestal niet groter dan duizend, maar dit aantal is voor het doel groot genoeg. Voorspellingen tijdens de verkiezingsdag zelf kunnen vanzelfsprekend nauwkeuriger, al naar gelang het aantal uitgebrachte stemmen.

1.3 Statistiek voor bedrijf en industriële productie

In de tweede helft van de vorige eeuw is de rol van de statistiek in het bedrijfsleven steeds groter geworden. Enerzijds is dat het gevolg van de steeds betere hulpmiddelen. Computers werden steeds sneller, waardoor ook grote verzamelingen gegevens snel geanalyseerd konden worden. De op het gebruik van statistiek gerichte computersoftware werd steeds krachtiger en gebruiksvriendelijker, onder andere door sterke verbeteringen in de grafische interface. Een voor velen herkenbaar voorbeeld is de televisieuitzending op de avond na de verkiezingen voor de Tweede Kamer.
Anderzijds is de behoefte aan het gebruik van statistiek sterk toegenomen, bijvoorbeeld ten gevolge van het streven naar steeds betere kwaliteit. In grotere bedrijven wordt statistiek tegenwoordig veelvuldig gebruikt in het kader van het totale kwaliteitsbeleid. Zeker in de westerse wereld (en in Japan) is statistische controle tijdens productieprocessen steeds belangrijker geworden. Aanvankelijk werden hiervoor kwaliteitscontroleurs ingeschakeld. In de laatste tien jaar heeft men in veel bedrijven kans gezien de procescontrole te automatiseren. Op basis van continue controle kan een computer zonder menselijke tussenkomst het productieproces bijsturen. Om te begrijpen wat er in dat geval gebeurt, is kennis van de statistiek noodzakelijk. Een geheel ander voorbeeld van het gebruik van statistiek in het bedrijfsleven ligt op het gebied van de marketing en de logistiek. Door statistisch onderzoek van marketinggegevens kan de behoefte aan bepaalde producten goed ingeschat worden.

Met behulp van statistische voorspelmethoden kan de vraag naar een bepaald product ook getalsmatig voorspeld worden. Diezelfde statistiek kan ook worden gebruikt om een efficiënte voorraadpolitiek te bereiken. Hierbij worden de vraag naar een product en de op te bouwen voorraad van dat product zodanig op elkaar afgestemd dat de kosten voor de producent zo laag mogelijk zijn.

1.4 Computer en statistiek

De tijd dat men statistiek bedreef met een zakrekenmachine is vrijwel voorbij. Wel beschikt vrijwel iedere eenvoudige 'scientific' pocketcalculator over enkele statistische functies. Maar zonder controle op de invoer van gegevens is zo'n apparaatje eigenlijk ongeschikt. Er is echter tegenwoordig veel computersoftware beschikbaar om waarnemingsuitkomsten te rangschikken, te karakteriseren en te presenteren. Ook zijn (vrijwel) alle wiskundige modellen die voor de toegepaste statistiek van belang zijn, in computersoftware beschikbaar. Spreadsheetprogramma's zoals EXCEL bieden talloze mogelijkheden. Nog meer geavanceerde toepassingen, ook ten aanzien van het invoeren van (soms massa's) gegevens, worden geboden door gespecialiseerde pakketten. We noemen SPSS, Minitab, SAS en ACTIVESTATS. In dit boek zullen we ons beperken tot een aantal toepassingen met behulp van EXCEL, omdat dit programma binnen ieders bereik ligt.

1.5 De fasen van een statistisch onderzoek

De vorm waarin verzamelde (en soms ook reeds bestaande) informatie ter beschikking komt, is meestal niet geschikt om daaruit rechtstreeks conclusies te trekken: men zal het eerst op een of andere wijze moeten ordenen. Om het risico op verkeerde conclusies en ten gevolge daarvan op foutieve beslissingen tot een minimum te beperken, dient een dergelijke ordening op verantwoorde wijze te geschieden. Ook nadat verzameld cijfermateriaal is geordend, zal het meestal nog niet mogelijk zijn er al conclusies uit te trekken: een verdere analyse van het verzamelde en geordende cijfermateriaal zal noodzakelijk zijn. Vaak wordt een dergelijke analyse voorafgegaan door het berekenen van allerlei karakteristieke grootheden van het cijfermateriaal, het zogenaamd karakteriseren.

1.5.1 Beschrijvende statistiek

Het is de *beschrijvende statistiek* die ons de hulpmiddelen biedt om cijfermateriaal te verzamelen, te ordenen en te karakteriseren. Vaak is de organisatie rondom het verzamelen, het ordenen, het karakteriseren en het analyseren van het voor een onderzoek benodigde cijfermateriaal zo complex, dat men de hulp van specialisten (statistici, marktonderzoekers, computerdeskundigen, vakspecialisten) moet inroepen. Zeker in dat geval – maar in feite altijd – zal men veel aandacht moeten besteden aan de voorbereiding van het onderzoek. Zo zal men vóór de aanvang van het onderzoek nauwkeurig moeten formuleren wat de

doelstelling van het onderzoek is en zal men moeten vaststellen welk cijfermateriaal in het kader van deze doelstelling van belang is en dus verzameld moet worden. Daarbij mag niet vergeten worden afspraken te maken over de te volgen waarnemingsmethode en over de daarbij te gebruiken hulpmiddelen en de te hanteren nauwkeurigheid. Verder dient men van tevoren vast te leggen welke analysemethoden gebruikt zullen worden. In de beschrijvende statistiek is het gebruik van een computer vanzelfsprekend geworden.

1.5.2 Toegepaste statistiek

Het onderdeel van de statistiek dat de methoden en technieken biedt om reeds verzameld, geordend en gekarakteriseerd cijfermateriaal te analyseren en om uit geanalyseerd cijfermateriaal conclusies te trekken, noemen we de *toegepaste statistiek*. Deze toegepaste statistiek (ook wel verklarende of analytische statistiek genoemd) is voor een belangrijk deel gebaseerd op de wetten van de kansrekening, reden waarom we ook wel spreken van stochastische statistiek (*stochas* = toeval). Aan de toegepaste statistiek liggen veel wiskundige modellen ten grondslag. In dit boek zullen we ons in algemene bewoordingen uitlaten over deze modellen. Het gaat in dit boek hoofdzakelijk om de toepassing van deze modellen en niet om de wiskundige verantwoording en de wijze waarop ze ontstaan. Op enkele wiskundige modellen zullen we wel wat dieper ingaan, omdat het gebruik ervan enig inzicht vereist.

1.6 Populatie en steekproef

Voor een statistisch onderzoek zijn altijd kwantitatieve gegevens nodig. Een kwantitatief gegeven is een gegeven waarin de waarde (bijvoorbeeld 178 cm) van een kenmerk (bijvoorbeeld lengte) van een object (bijvoorbeeld een man van 21 jaar) in de vorm van een getal wordt vastgelegd. De vaststelling van de waarde van een kenmerk van een object noemen we een *waarneming*, de vastgestelde waarde een *waarnemingsuitkomst*. Behalve de doelstelling van een statistisch onderzoek dient ook de te onderzoeken *populatie* zo nauwkeurig mogelijk omschreven te worden.

Definitie
Onder een *populatie* verstaat men de verzameling van alle kwantitatieve gegevens die in het kader van de doelstelling van een statistisch onderzoek van belang zijn en die dan de elementen van de verzameling worden genoemd.

In een onderzoek naar de lengte van de Nederlandse man van 21 jaar kan de populatie omschreven worden als de verzameling lengten van *alle* Nederlandse mannen van 21 jaar.
Een populatie bezit bepaalde karakteristieken, waarvan sommige in een getal zijn vast te leggen. Een karakteristieke grootheid van een populatie noemen we een *parameter* van de populatie. In het onderzoek naar de lengte van de Nederlandse man van 21 jaar is de

gemiddelde lengte een voor de hand liggende parameter. Maar de populatie van lengten bevat meer parameters, zoals we later zullen zien.
Een volgend aspect waaraan men bij het voorbereiden van een statistisch onderzoek aandacht dient te besteden, is de aard van het onderzoek: 100%-onderzoek of *steekproefsgewijs* onderzoek. Populaties zijn vaak zeer groot en soms zelfs 'oneindig' groot. In het eerste geval is een 100%-onderzoek van de totale populatie vaak te kostbaar en in het tweede geval zelfs onmogelijk. Maar ook bij een eindige populatie kan een 100%-onderzoek onmogelijk zijn, bijvoorbeeld bij een destructief onderzoek waarbij de objecten van onderzoek ter wille van het onderzoek vernietigd moeten worden (onderzoek op breekbaarheid, levensduur, ontvlambaarheid, enzovoorts). In deze en dergelijke gevallen volstaat men met een *steekproef* uit de populatie.

Definitie
Een *steekproef* uit een populatie is een deelverzameling van die populatie en bevat een eindig aantal waarnemingsuitkomsten.

In het kader van het eerder genoemde onderzoek vormen de lengten van alle Amsterdamse mannen van 21 jaar een steekproef uit de gedefinieerde populatie. Of het statistisch gezien een goede steekproef is, zou men kunnen betwijfelen. Men dient zich namelijk te realiseren dat aan het werken met steekproeven bepaalde risico's verbonden zijn: allerlei karakteristieke grootheden van een steekproef, zoals de gemiddelde lengte, zijn slechts een *schatting* (officiële benaming: *schatter*) van de overeenkomstige parameter (in dit geval: gemiddelde lengte) van de populatie. De resultaten van een steekproefonderzoek bezitten daardoor een zekere mate van onnauwkeurigheid [1] en een zekere mate van onbetrouwbaarheid, waardoor het risico ontstaat dat men verkeerde conclusies trekt en daardoor verkeerde beslissingen neemt. Men dient dus de nodige voorzichtigheid te betrachten bij het nemen van steekproeven. We komen er in de volgende hoofdstukken op terug.

1.7 Steekproefonderzoek (voorbeschouwing)

Hierboven is al aangegeven wat het verschil is tussen een populatie en een steekproef. In deze paragraaf zullen we een voorbeschouwing maken waarin we enkele vragen zullen opwerpen over de wijze waarop een steekproef genomen moet worden.
Een eerste punt van belang bij het voorbereiden van een statistisch onderzoek is dat men – of men nu kiest voor een populatie-onderzoek of voor een steekproefonderzoek – vóór de aanvang van het onderzoek vaststelt volgens welke methode (marktonderzoek, enquête, simulatie, kwaliteitsonderzoek, enzovoorts) het onderzoek zal plaatsvinden. Daarbij komen

[1] Op de begrippen (on)nauwkeurigheid en (on)betrouwbaarheid komen we later in het boek uitvoerig terug. Thans zij slechts provisorisch opgemerkt dat de conclusie "de gemiddelde lengte ligt tussen 170 en 180 cm" onnauwkeuriger is dan de conclusie "de gemiddelde lengte ligt tussen 174 en 176 cm". De onbetrouwbaarheid van dit soort conclusies is de kans dat zij niet juist zijn.

vragen aan de orde als: Hoe en met welke hulpmiddelen zullen de waarnemingen worden verricht? Hoe en met welke nauwkeurigheid zullen de waarnemingsuitkomsten worden vastgelegd? Hoe zullen de data worden verwerkt? Geeft de te volgen methode betrouwbaar en nauwkeurig genoeg de gewenste en/of noodzakelijke informatie? Wegen de kosten van de te volgen methode op tegen het voordeel dat men door het nemen van een juiste beslissing (of juist het vermijden van een onjuiste beslissing!) hoopt te bereiken?
Wanneer de waarnemingsuitkomsten zijn verzameld, gaat men er, vrijwel altijd met behulp van een computerprogramma, toe over deze te ordenen. Dit houdt in dat men de verzamelde waarnemingsuitkomsten gaat sorteren (in een gewenste volgorde gaat rangschikken) en/of gaat presenteren in de vorm van tabellen en/of grafieken (diagrammen). Nadat de verzamelde data geordend en/of gepresenteerd zijn, worden de benodigde en/of gewenste *karakteristieke grootheden* ervan berekend. Karakteristieke grootheden zijn getallen waarmee men alle waarnemingsuitkomsten als het ware kan karakteriseren of samenvatten. Een voorbeeld is het (rekenkundig) gemiddelde, dat uiteraard alleen bruikbaar is bij kwantitatieve variabelen (dit zijn variabelen waarvan de meetwaarde een getal is).
Met het verzamelen, ordenen, presenteren en karakteriseren is de fase van de beschrijvende statistiek voltooid. Nu volgt de fase van de toegepaste statistiek: de verdere analyse van de waarnemingsuitkomsten, waarvan moet afhangen welke conclusies de uiteindelijk te nemen beslissingen moeten ondersteunen.

1.7.1 Het nemen van steekproeven

De belangrijkste vraag bij het nemen van een steekproef is hoe deze uit alle denkbare data van de populatie moet worden samengesteld. De bedoeling daarbij is om tegen minimale kosten met voldoende nauwkeurigheid en voldoende betrouwbaarheid (nogmaals: de begrippen nauwkeurigheid en betrouwbaarheid zijn niet hetzelfde en zullen later worden uitgelegd) conclusies te kunnen trekken over de karakteristieke grootheden van de gehele populatie. Om dit te kunnen realiseren, zal de steekproef een goede afspiegeling moeten zijn van de populatie, dat wil zeggen zal de steekproef *representatief* moeten zijn (re-presenteren = opnieuw voorkomen).

Definitie

Een steekproef heet *representatief* wanneer alle (denkbare) eigenschappen van het te onderzoeken kenmerk in de populatie in voldoende mate in de steekproef vertegenwoordigd zijn.

Voorbeeld: men wil een enquête houden over het kiezersgedrag, voorafgaande aan de verkiezingen voor de Tweede Kamer. De populatie bestaat uit alle stemgerechtigde personen met een Nederlandse nationaliteit. Wat is een representatieve steekproef? Deze vraag is niet eenvoudig te beantwoorden. Een steekproef zal pas representatief zijn wanneer daarin een acceptabele verhouding bestaat tussen het aantal jongste kiezers (zeg onder 25 jaar), het aantal jonge kiezers (zeg tussen de 25 en de 45), het aantal kiezers van middelbare leeftijd

(zeg tussen 45 en 60 jaar) en het aantal kiezers dat ouder is. Daarmee zijn we er echter niet. De steekproef zal ook een afspiegeling moeten zijn tussen het aantal Nederlandse mannen en vrouwen. En tussen het aantal allochtone en autochtone Nederlanders. Er zal bovendien een goede afspiegeling moeten zijn tussen het aantal inwoners in grote steden en het aantal Nederlanders dat juist niet in gote steden woont. En ga zo maar door... In zijn algemeenheid zal men bij de beoordeling van de representativiteit van zo'n steekproef moeten nagaan welke eigenschappen van de te ondervragen personen in de populatie van belang zouden kunnen zijn (bijvoorbeeld leeftijd, geslacht, burgerlijke staat, opleidingsniveau, godsdienstige en/of politieke overtuiging, woonplaats) en vervolgens moeten nagaan in hoeverre deze eigenschappen in de steekproef vertegenwoordigd zijn. Een steekproef met relatief te veel of te weinig ouderen, mannen, alleenwonenden, academici, rooms-katholieken, sociaal-democraten of Hagenaars zal in dat geval niet of niet voldoende representatief zijn. Een goed hulpmiddel om te bereiken dat een steekproef representatief genoeg is, vinden we in het nemen van een *aselecte* steekproef (a-select = niet uitgezocht).

Definitie

Een steekproef van n waarnemingsuitkomsten uit een populatie van N waarnemingsuitkomsten heet *aselect* wanneer elke deelverzameling van n waarnemingsuitkomsten uit de N waarnemingsuitkomsten van de populatie gelijke kans heeft om de te nemen steekproef te vormen.

Merk op dat de begrippen representatief en aselect wel met elkaar samenhangen maar niet dezelfde betekenis hebben. Bij een steekproef die niet aselect is genomen, bestaat het gevaar dat deze niet representatief is. Maar een aselecte, doch ten opzichte van de omvang van de populatie relatief te kleine steekproef, zal meestal niet voldoende representatief zijn. Men kan bereiken dat een steekproef aselect is door ervoor te zorgen dat elk element in de populatie gelijke kans heeft om in de steekproef te worden opgenomen. Dit kan men bereiken door bij het kiezen van een waarnemingsuitkomst uit de populatie geen enkele persoonlijke voorkeur te laten gelden voor het waar te nemen kenmerk. Of anders gezegd: door bij het verrichten van waarnemingen geen enkele vorm van subjectiviteit te laten meespelen.
Het verdient aanbeveling daarvoor een methode te gebruiken waarbij door middel van *loting* van tevoren wordt vastgesteld welke data uit de populatie in de steekproef zullen worden opgenomen. Als hulpmiddel daarbij kan men gebruikmaken van lotingstabellen of van een randomgenerator. Een op dergelijke wijze genomen steekproef wordt een *gelote steekproef* genoemd (Engels: *random*).
Zoals reeds eerder is gezegd, neemt men uit een populatie een steekproef omdat het te kostbaar is (bij grote populaties) of omdat het onmogelijk is (bij oneindig grote populaties of bij destructief onderzoek) om de gehele populatie te onderzoeken. Dit betekent niet dat men dus per definitie altijd een steekproef neemt. Zo zal men bijvoorbeeld bij kleine populaties vaak de gehele populatie onderzoeken (mits het onderzoek niet destructief is). Dat zal men ook doen, zelfs moeten doen, wanneer men niet het risico kan lopen op verkeerde conclu-

sies en dus verkeerde beslissingen. Dit speelt bijvoorbeeld een rol bij het ontwerpen en het vervaardigen van producten die bij verkeerd functioneren een bedreiging kunnen vormen voor het welzijn, de gezondheid of misschien zelfs het leven van de gebruiker. Denk hierbij aan levensmiddelen, geneesmiddelen, smaak- en voedingsstoffen, cosmetica, gebruiksvoorwerpen, elektrische apparaten, vervoersmiddelen, enzovoorts. In dat kader mogen we van geluk spreken dat het, met name in de industrie, maar ook in andere toepassingsgebieden, als gevolg van de steeds verder voortschrijdende automatisering en mechanisering steeds eenvoudiger en minder kostbaar wordt om 100%-onderzoek te doen in plaats van steekproefonderzoek.

1.7.2 Het organiseren van enquêtes

Een veelgebruikte methode van statistisch onderzoek vinden we in de *enquête*. Als afsluiting van dit hoofdstuk willen we kort ingaan op het nemen van enquêtes.

Enquêtes worden meestal uitgevoerd door gespecialiseerde instanties en onderzoekbureaus omdat de voorbereiding en de organisatie ervan erg zorgvuldig dient te geschieden, waarvoor deskundigheid een eerste vereiste is.

In principe kan een enquête op twee verschillende manieren worden uitgevoerd: mondeling (al of niet telefonisch) en schriftelijk. Elk van beide methoden bezit zijn eigen voor- en nadelen.

Een schriftelijke enquête kost in de uitvoeringsfase alleen portokosten en men kan er relatief veel mensen in een betrekkelijk korte tijd mee ondervragen. Een mondelinge enquête daarentegen is in de uitvoeringsfase zeer arbeidsintensief (hoge loonkosten) en kost veel tijd en/of geld. Daar staat tegenover dat een schriftelijke enquête doorgaans een hoog percentage non-response – het percentage van de ondervraagden dat om welke reden dan ook het enquêteformulier niet terugstuurt – kent, terwijl de meeste mensen die mondeling (al of niet telefonisch) door een enquêteur worden benaderd doorgaans meer bereid zijn de gestelde vragen te beantwoorden.

Het optreden van het non-response-verschijnsel bij schriftelijke enquêtes is enigszins te beperken door de ondervraagden een beloning in het vooruitzicht te stellen (hetgeen echter weer kostenverhogend werkt). Men loopt echter het risico dat de uiteindelijk resterende steekproef van degenen die wel reageren niet meer representatief is. Bij een mondelinge enquête loopt men weer het risico dat door de veelheid van enquêteurs (overigens enigszins te beperken door de enquête telefonisch te houden) de eenduidigheid van de gestelde vragen en de objectiviteit van de gegeven antwoorden in gevaar komt.

Het organiseren van enquêtes is niet de taak van de statisticus alleen. Zeker in de voorbereidingsfase is dit in veel gevallen een kwestie van gemeenschappelijke zorg van een statisticus, een gedragswetenschapper (socioloog of psycholoog) en iemand die goed kan omgaan met daarvoor geschikte computerprogrammatuur.

2 Het verzamelen van data

2.1 Inleiding

Aan elk statistisch onderzoek gaat het verzamelen van gegevens, waarnemingen of waarnemingsresultaten (samengevat in het woord *data*) vooraf. In dit hoofdstuk zullen we uitvoerig daarop ingaan. We zullen het hebben over de wijze waarop data verzameld worden, hoe ze geordend worden en hoe ze gepresenteerd kunnen worden. In de statistiek is het van groot belang om data naar soort te kunnen onderscheiden. Daartoe moeten we eerst het begrip variabele nauwkeurig definiëren. Het begrip variabele kennen we uit de wiskunde als een grootheid (bijvoorbeeld x of y) waaraan een getalswaarde (al of niet in een bepaalde eenheid) toegekend kan worden. In de statistiek is het begrip variabele veel ruimer gedefinieerd.

2.2 Het begrip variabele

Definitie

Wanneer een 'kenmerk' van een 'object' bij waarneming aan meerdere exemplaren van dat object niet noodzakelijkerwijs steeds dezelfde waarde oplevert, zegt men dat het kenmerk variabel is en noemt men het een *variabele*.

Een kenmerk van een object als bedoeld in bovenstaande definitie is bijvoorbeeld het gewicht van een pasgeboren kind, de leeftijd van een kat, het vitaminegehalte van een tomaat of de levensduur van een gloeilamp. Maar ook: het aantal kinderen in een gezin, het aantal honden in een woonwijk, het aantal sinaasappelen in een krat of het aantal defecte exemplaren in een dagproductie. Of: de hoeveelheid neerslag per dag, de hoeveelheid rode bloedlichaampjes per centiliter bloed of de hoeveelheid zuurstofatomen per kubieke centimeter lucht. Vaak kan de waarde van een kenmerk van een object (dus de waarde van een variabele) in een reëel getal – de waarnemingsuitkomst – worden vastgelegd, maar dit is niet altijd het geval. Om tussen deze beide typen variabelen onderscheid te maken, definieert men:

Definitie
Wanneer de waarde van een kenmerk van een object in een reëel getal kan worden uitgedrukt, noemt men dat kenmerk een *kwantitatieve* variabele.

Definitie
Wanneer de waarde van een kenmerk van een object niet in een reëel getal kan worden uitgedrukt, noemt men dat kenmerk een *kwalitatieve* variabele.

Kwantitatieve variabelen worden onderscheiden in continue variabelen (ook wel meetbare grootheden genoemd) en discrete of discontinue variabelen (ook wel niet-meetbare of telbare grootheden genoemd).

Definitie
Een *continue* variabele is een kwantitatieve variabele waarvan de waarde kan worden uitgedrukt in elk reëel getal op een zeker interval.

Definitie
Een *discrete* (of discontinue) variabele is een kwantitatieve variabele die op een zeker interval slechts bepaalde waarden (meestal natuurlijke getallen) kan aannemen.

Kwalitatieve variabelen worden onderscheiden in variabelen die *rangschikbaar* zijn (bijvoorbeeld de kleuren van de regenboog of de rangen van militairen) en variabelen die niet-rangschikbaar zijn (bijvoorbeeld de namen van politieke partijen of de merknamen van computerapparatuur). Rangschikbare kwalitatieve variabelen worden weer onderscheiden in continu-rangschikbare variabelen (de kleuren van de regenboog) en discreet-rangschikbare variabelen (de rangen van militairen).

Opdracht
Beschrijf voor de volgende voorbeelden de in bovenstaande definities gehanteerde begrippen 'object' en 'kenmerk' en ga na tot welk type (kwalitatief - kwantitatief, continu - discreet) de variabelen behoren.
- De kleur van de ogen van een pasgeboren baby.
- De stroomsterkte in een elektrisch netwerk.
- Het aantal pinda's in zakjes van 100 gram.
- Het aantal geboren pandabeertjes per jaar.
- De treksterkte van betonstaal van 1 cm dikte.
- De mate van waardering voor een bepaalde politicus ('goed', 'matig', 'slecht').
- De omzet per dag van een autoverhuurbedrijf.
- De kijkdichtheid van het NOS-journaal.
- De hoeveelheid microben in een kubieke centimeter slootwater.
- Het aantal kinderen in een gezin.

- De nationaliteit van in Nederland woonachtige buitenlanders.
- Het totaal aantal bioscoopbezoekers van een bepaalde film.
- De economische levensduur van een auto.
- Het aantal dagen per jaar met meer dan 5 mm neerslag in De Bilt.
- Het gemiddeld benzinegebruik per 100 km van een auto.

Wanneer het onzeker is, dat wil zeggen wanneer het van het toeval afhangt welke waarde een kwantitatieve variabele bij waarneming zal aannemen (met andere woorden: welke waarnemingsuitkomst men zal vinden), noemt men die variabele een kansvariabele. Dus:

Definitie

Een *kansvariabele* (ook wel stochast genoemd, *stochas* = toeval) is een continue of discrete kwantitatieve variabele waarvan het van het toeval afhangt welke waarde deze bij waarneming zal aannemen.

De naam van een kansvariabele wordt meestal aangeduid met een hoofdletter (bijvoorbeeld U, V, W), eventueel voorzien van een index (bijvoorbeeld X_1, Y_1, Z_1). De waarde ervan (de waarnemingsuitkomst) wordt – zo deze nog niet bekend is – aangeduid met de overeenkomstige kleine letter (dus u, v, w, x_1, y_1, z_1). Dit houdt in dat bijvoorbeeld de op het eerste gezicht vreemde notatie $X = x$ mogelijk is: de variabele met de naam X heeft als waarde x.

Wanneer y volgens het een of andere functievoorschrift $y = f(x)$ niet van het toeval afhangt maar afhangt van de gekozen waarde x van een ander kenmerk X, dan heet kenmerk Y (met waarde y) een *deterministische* variabele. Dit is ook het geval wanneer de waarde y bij elke waarneming aan Y hetzelfde is. Kenmerk Y is dan niet variabel maar deterministisch, in het laatste geval zelfs constant. Zo is bijvoorbeeld het jaarlijks te betalen rentebedrag Y op een hypotheek (mits de rente constant is) een deterministische variabele. De waarde y van dit bedrag is namelijk niet van het toeval afhankelijk, maar is – althans bij gelijkblijvende rentevoet – voor een levensverzekeringshypotheek constant en voor een annuïteit of een lineaire hypotheek volgens een bepaald functievoorschrift $y = f(x)$ afhankelijk van de waarde x van het nog af te lossen bedrag X.

2.3 Het meetniveau van een variabele

Nadat aan een variabele metingen verricht zijn, zal men in veel gevallen de meetresultaten willen bewerken, bijvoorbeeld om het gemiddelde ervan te berekenen. Rekenkundige bewerkingen zijn echter lang niet altijd mogelijk. Het is natuurlijk onzin om de gemiddelde kleur ogen van 100 blanke kinderen te bepalen. Maar ook het middelen van hun geboortejaren is discutabel. Daarentegen kan het wel zinvol zijn om het gemiddelde te bepalen van hun leeftijden. Om te kunnen bepalen of rekenkundige bewerkingen wel of niet mogelijk zijn, worden variabelen gesorteerd naar hun *meetniveau*. Bij elk meetniveau wordt een bepaalde

schaal gehanteerd. Er zijn vier schalen: de *ordinale* en *nominale schaal* voor kwalitatieve variabelen; de *ratio-* en de *intervalschaal* voor kwantitatieve variabelen.

2.3.1 Nominale schaal en ordinale schaal

Wanneer een kwalitatieve variabele op geen enkele zinvolle wijze in een bepaalde volgorde gerangschikt kan worden, is een *nominale* schaal noodzakelijk. De namen van politieke partijen (PvdA, CDA, VVD, enzovoorts), de kleur van ogen (blauw, bruin, groen, enzovoorts), het geslacht (M/V), het feit of men wel eens een stickie gerookt heeft of niet (Ja/Nee); dit zijn variabelen die op een nominale schaal moeten worden aangebracht. Voor kwalitatieve variabelen waarvoor het wel zinvol is ze in een bepaalde volgorde te ordenen, wordt een *ordinale* schaal gebruikt. Voorbeelden: de smaak van een bepaald merk soep (lekker, matig, niet lekker), de rang van een militair (korporaal, sergeant, luitenant, enzovoorts), de kwaliteit van een product (−−, −, 0, +, ++, dan wel: zeer slecht, slecht, redelijk, voldoende, goed). Hoewel een ordinale schaal iets beter weergeeft wat het verschil in de uitkomst van de variabele is, blijft het een zwakke vorm van meting. De waarde van het verschil tussen twee waarnemingsuitkomsten kan niet bepaald worden.

We merken nog op dat het vaak voorkomt (vooral bij gebruik van een computerprogramma) dat de uitkomsten van zowel variabelen met een nominale schaal als variabelen met een ordinale schaal vervangen worden door een natuurlijk getal. Voorbeeld: Opel = 1, Volkswagen = 2, Audi = 3 (variabele automerk, nominaal) of 'zeer slecht' = 1, 'slecht' = 2, 'matig' = 3, 'goed' = 4, 'uitstekend' = 5 (kwaliteitsoordeel, ordinaal). We moeten daarbij wel bedenken dat het schaalkarakter door het coderen niet verandert, al lijkt dat wel zo! De volgorde (bij een nominale schaal) van, dan wel het verschil (bij een ordinale schaal) tussen de uitkomsten blijft immers ook na het coderen willekeurig.

2.3.2 Intervalschaal en ratioschaal

Op een *intervalschaal* is het verschil tussen twee waarnemingsuitkomsten meetbaar en van betekenis. Bijvoorbeeld, een man met een lengte van 1,78 m is 10 cm langer dan een man met een lengte van 1,68 m. Dat verschil van 10 cm heeft dan weer dezelfde betekenis wanneer twee mannen met lengten van 1,96 m en 1,86 m vergeleken zouden worden. Tijden (jaren, maanden, dagen) zijn ook variabelen waar een intervalschaal van toepassing is. Immers het verschil tussen twee verschillende tijden is meetbaar en van betekenis. Het spreekt vanzelf dat een intervalschaal alleen van toepassing is op kwantitatieve variabelen

Bij een *ratioschaal* (ook uitsluitend voor kwantitatieve variabelen) is het meetniveau nog hoger. Daarbij is ook een natuurlijk nulpunt aanwezig dat het mogelijk en vaak ook zinvol maakt verschillende waarnemingsuitkomsten op elkaar te delen. Zo kan men stellen dat iemand met een gewicht van 100 kg twee keer zo zwaar is als iemand met een gewicht van 50 kg. Voor de variabele 'lengte' kan men dus ook een ratioschaal gebruiken. Maar voor de variabele 'tijd' is uitsluitend een intervalschaal bruikbaar. Het jaar 2000 (als historisch meetpunt) is uiteraard niet tweemaal zo groot als het jaar 1000.

Opdracht

Geef aan van welke soort de volgende variabelen zijn en welk meetniveau ze hebben.

- fruitsoort (appel, peer, sinaasappel, meloen);
- tentamencijfer (1, 2, 3, ..., 10);
- gemiddelde dagtemperatuur (gemeten in graden Celsius);
- absolute temperatuur (gemeten in graden Kelvin);
- aantal hartslagen per minuut;
- treksterkte van een staaf;
- kleur;
- gewichtsklasse (bijvoorbeeld van een bokser);
- reistijd;
- geboortejaar.

2.4 Het samenstellen van tabellen en het tekenen van grafieken

Naast het verzamelen van statistische gegevens (waarnemingsuitkomsten) behoort het ook tot het werkterrein van de beschrijvende statistiek deze gegevens op overzichtelijke wijze te ordenen in tabellen en grafieken. In deze paragraaf zullen we aan dit aspect van de beschrijvende statistiek enige aandacht besteden.

2.4.1 Het samenstellen van tabellen

Om de waarnemingsuitkomsten van een statistisch onderzoek in de vorm van een tabel te presenteren, moet deze er aantrekkelijk uitzien. Dit kan men onder andere bereiken door ervoor te zorgen dat de tabel is voorzien van een duidelijk en volledig op- of onderschrift en van een korte maar duidelijke omschrijving van de betekenis van de gebruikte regels en kolommen (bijschriften). De leesbaarheid van een tabel wordt in hoge mate bevorderd door een logische en overzichtelijke indeling. In het algemeen bestaat bij het samenstellen van tabellen de gewoonte liever enige nauwkeurigheid op te offeren wanneer daarmee bereikt kan worden dat de leesbaarheid wordt bevorderd. Zo zal men numerieke gegevens bij voorkeur presenteren in niet meer dan 4 cijfers, eventueel – na afronding – in (bij voorkeur drievoudige) machten van 10. Zo zal bijvoorbeeld het getal 356183 gepubliceerd worden als $356{,}2 \times 10^3$ of als $0{,}356 \times 10^6$. Tabel 2.1 geeft een voorbeeld van tabelletje van 5 objecten met verschillende meetkenmerken.

Tabel 2.1 Lengte van 5 studenten

nummer	geslacht	leeftijd	lengte (in m)
1	m	18	1,78
2	v	19	1,69
3	m	17	1,60
4	v	18	1,78
5	m	18	1,80

In tabel 2.2 is in één oogopslag te zien hoe de waarnemingsuitkomsten over het gemeten interval verdeeld zijn.

Tabel 2.2 Frequentietabel van de lengte van 25 vrouwen en 25 mannen

lengte (in m)	geslacht		totaal
	m	v	
1,65-<1,70	1	4	5
1,70-<1,75	4	10	14
1,75-<1,80	10	6	16
1,80-<1,85	6	4	10
1,85-<1,90	4	1	5
totaal	25	25	50

Statistische software zoals EXCEL bevat vele mogelijkheden om naar eigen wens een tabel te presenteren. Via internet kunnen van vele verschijnselen tabellen worden bekeken en gedownload. Om kerngegevens over economische verschijnselen te verzamelen, is een bezoek op de website van het Centraal Bureau voor de Statistiek (www.cbs.nl) van harte aan te bevelen.

2.4.2 Grafieken en afbeeldingen

Een andere manier om verzamelde gegevens op overzichtelijke wijze te rangschikken en te presenteren, vinden we in het tekenen van grafieken, ook wel diagrammen genoemd. Het gedrag van statistisch onderzochte verschijnselen komt in een grafiek beter tot zijn recht. Eventueel bestaande samenhang tussen verschillende variabelen kan in een grafiek duidelijker herkend worden dan in een tabel. Er bestaan verschillende soorten grafieken, waarvan we enkele voorbeelden laten zien. Alle voorbeelden betreffen min of meer fictieve waarnemingen. Zij zijn gemaakt in EXCEL, dat een zeer krachtige ondersteuning biedt voor het maken van vele soorten grafieken en afbeeldingen.

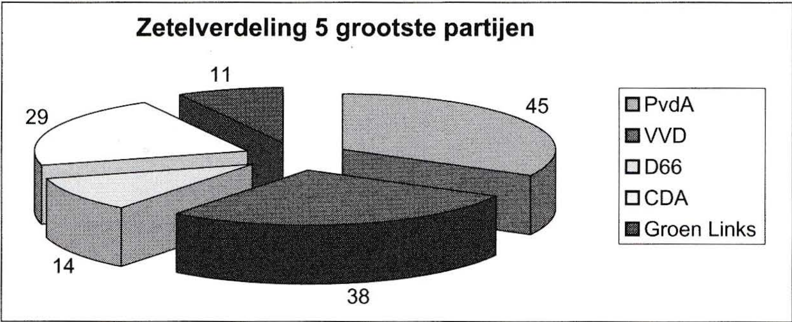


Fig. 2.1 Zetelverdeling Tweede kamer 2001

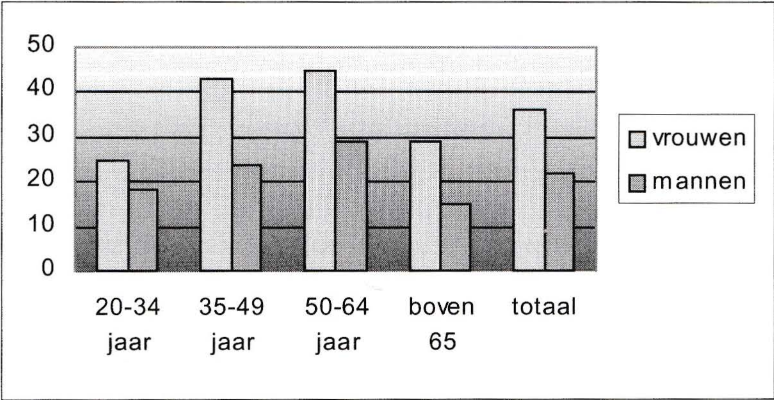


Fig. 2.2 Percentage mannen en vrouwen dat zichzelf te dik vindt

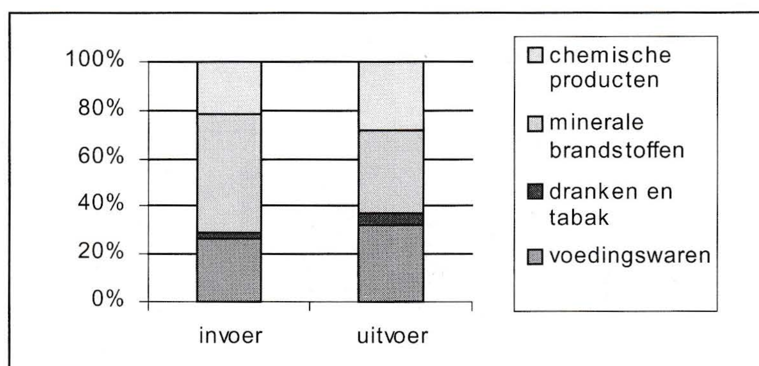


Fig. 2.3 In- en uitvoer van enkele categorieën producten

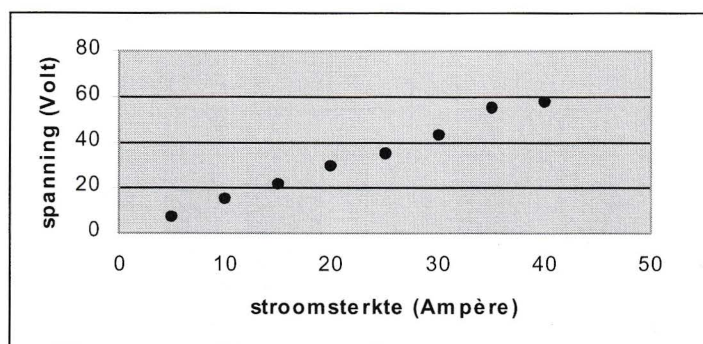


Fig. 2.4 Wet van Ohm

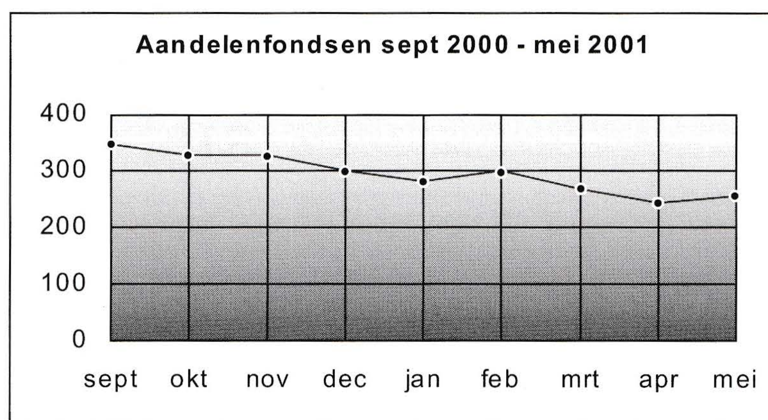


Fig. 2.5 Verloop aandelenfondsen

In bijlage A zullen we globaal laten zien hoe met EXCEL grafieken en afbeeldingen gemaakt kunnen worden.
Bij het tekenen van grafieken dienen we erop te letten dat de figuur voorzien is van een op- of onderschrift en van bijschriften langs de beide assen. Indien de grafieken niet op basis van eigen waarnemingen zijn gemaakt, dient een bronvermelding te worden gegeven.
In figuur 2.1 zien we een *cirkeldiagram* van het aantal zetels in de Tweede Kamer voor de vijf grootste politieke partijen. Zo'n diagram wordt gebruikt om snel de verschillen te kunnen laten zien. In figuur 2.2 zien we een *kolom- of staafdiagram* waarin twee groepen met elkaar vergeleken worden. Figuur 2.3 is een *staafstapeldiagram* waarin vier groepen op twee verschillende manieren met elkaar vergeleken kunnen worden. Wil men van een of meerdere kwalitatieve of kwantitatieve variabelen laten zien met welke frequentie deze in verschillende klassen voorkomen, dan gebruikt men een (samengesteld) kolommendiagram of een staafstapeldiagram. In figuur 2.4 zien we een *puntendiagram* waarin meetresultaten tweedimensionaal tegen elkaar worden afgezet. Puntendiagrammen (ook wel *scatterdiagrammen* genoemd) – zie figuur 2.4 – worden gebruikt om te onderzoeken of er tussen twee kwantitatieve variabelen een bepaalde samenhang bestaat respectievelijk om te laten zien dat een dergelijke samenhang bestaat.
In figuur 2.5 zien we een *lijndiagram* (ook wel polygoon genoemd). Hier is het gebruikt om een zogenaamde *tijdreeks* weer te geven. Een tijdreeks is een aantal waarnemingen die gedaan zijn na gelijke tijdsintervallen.
Er zijn nog veel meer soorten grafieken of diagrammen mogelijk. We hoeven de krant er maar op na te slaan om voorbeelden aan te treffen. In EXCEL is het aantal variaties van grafieken eveneens zeer groot.
In het volgende hoofdstuk zullen we nog het *histogram* tegenkomen.

3 Het weergeven en karakteriseren van data

3.1 Inleiding

In dit hoofdstuk wordt de kern behandeld van de beschrijvende statistiek. De waarnemingsresultaten, meetwaarden oftewel *data* die bij het statistisch onderzoek (populatie of steekproef) betrokken worden, zullen vrijwel altijd in klassen (of categorieën) worden verdeeld. We spreken dan van *frequentieverdelingen*.

3.2 Frequentieverdelingen

Het verdelen in klassen begint met het opstellen van een frequentietabel.

3.2.1 Frequentietabel

Een groot aantal waarnemingen geeft een onoverzichtelijk geheel, als er geen ordening is toegepast. Om een beter inzicht in de getallenmassa te krijgen, geeft men de gegevens weer in een overzichtelijke tabellen (= *frequentietabellen*). Aan de hand van een voorbeeld gaan we dit nader toelichten.

Voorbeeld 1

We hebben de beschikking over de uitkomsten van een steekproef van 50 gewichtsmetingen van een afvulmachine (in grammen, afgerond op een geheel getal). De gegevens zijn vermeld in tabel 3.1.

Opmerking

De data van voorbeeld 1 zijn gehele getallen. De gemeten variabele (gewicht van een afvulling) zal dan ook opgevat worden als een discrete variabele. We moeten ons echter goed realiseren dat een gewicht in principe een reëel getal is. Daarom hebben we het eigenlijk over een continue variabele. Dat we de gewichten toch beschouwen als waarden van een discrete variabele, komt uitsluitend omdat we de data hebben afgerond op gehele getallen.

Tabel 3.1 Resultaten van 50 gewichtsmetingen (in grammen)

96	91	94	106	89	89	94	94	95	97
99	87	87	96	96	96	86	81	96	96
91	91	92	96	99	99	99	96	92	88
88	95	87	94	95	95	103	90	92	111
84	94	110	96	88	88	96	96	103	92

Daar de gegevens van tabel 3.1 in volgorde van de metingen zijn gegeven, geeft dit geen duidelijk beeld van de getallenmassa. Om een beter inzicht te krijgen, zetten we de gegevens in volgorde van grootte en noteren hoe vaak een bepaalde uitkomst voorkomt. Dit is het basisprincipe van een frequentietabel. De resultaten staan weergegeven in tabel 3.2.

Tabel 3.2 Resultaten van 50 gewichtsmetingen

meetwaarde	aantal	meetwaarde	aantal	meetwaarde	aantal	meetwaarde	aantal
80	-	90	2	100	-	110	1
81	1	91	4	101	-	111	1
82	-	92	4	102	2	112	-
83	-	93	1	103	-	113	-
84	1	94	6	104	-	114	-
85	-	95	3	105	-	115	-
86	1	96	10	106	1	116	-
87	3	97	1	107	-	117	-
88	3	98	-	108	-	118	-
89	2	99	3	109	-	119	-

Tabel 3.2 geeft een duidelijker beeld van de verdeling van de verschillende meetuitkomsten. Een nog duidelijker beeld ontstaat als de gegevens van de tabel worden weergegeven in een grafiek. In figuur 3.1 is dit uitgevoerd. De frequentie per meetuitkomst is in kolomvorm boven de daarbij behorende gewichtswaarden uitgezet. De hoogte van de kolom correspondeert met de frequentie per meetuitkomst De kolommen liggen in principe tegen elkaar, in tegenstelling tot een kolom- of staafdiagram (waarin de staven los van elkaar opgericht zijn). Een dergelijke grafiek noemen we een *histogram.*

Ondanks het feit dat figuur 3.1 een goed overzicht geeft van de verdeling van meetuitkomsten van de steekproef, geeft deze nog geen goede 'beschrijving' van de werkelijkheid (=populatie), er zijn te veel lege 'plekken' in de steekproefverdeling, die bij de populatieverdeling vermoedelijk niet zullen voorkomen.
Om een goede beschrijving van de werkelijkheid te krijgen op grond van steekproefuitkomsten, moeten we een andere procedure volgen. Hiervoor is een standaardprocedure (ISO-norm) opgesteld, die we in de volgende subparagraaf zullen bespreken.

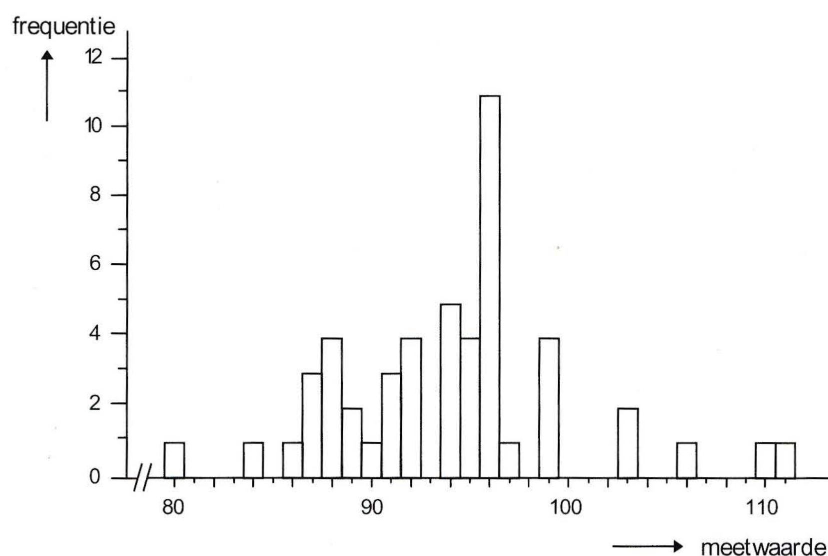


Fig. 3.1 Histogram van 50 gewichtsbepalingen (Tabel 3.2, klassenbreedte b = 1)

3.2.2 Het opstellen van een frequentietabel

Om een beter inzicht te krijgen in de verdeling van de populatie zijn we niet zo zeer geïnteresseerd in de aantallen per meetuitkomst, maar meer in het aantal per voorgeschreven interval. Als we bijvoorbeeld de lengte (in cm) meten van een aantal personen, dan zijn de mogelijke uitkomsten getallen tussen 50 en 250. Om een indruk van de verdeling van de populatie (= totale bevolking) te krijgen, zijn we niet zozeer geïnteresseerd in elke waarneming afzonderlijk, als wel in de 'frequentiedichtheid', dat wil zeggen we willen graag weten hoeveel waarnemingen er in een bepaald interval liggen. In de bovengenoemde lengtemeting interesseert het ons minder hoeveel mensen in de steekproef een lengte hebben van bijvoorbeeld 154 cm of 157 cm. Belangrijker is het te weten dat er van de bijvoorbeeld 100 mensen in de steekproef er 5 een lengte hebben tussen 150 en 160 cm en 10 een lengte hebben tussen 160 en 170 cm. Hierdoor krijgen we een veel beter beeld van de populatie, waaruit de steekproef is getrokken. Dit is in principe de basisregel van een statistisch onderzoek. De steekproefresultaten op zichzelf zijn niet de hoofdzaak van een onderzoek. De steekproefresulaten hebben we nodig om een beeld te krijgen van de werkelijkheid.

Nu terug naar ons voorbeeld met de gegevens van tabel 3.1. We verdelen het totale meetinterval, waarbinnen de waarnemingsuitkomsten vallen, in een aantal kleinere intervallen, meestal *klassen* genoemd. Het begin- en eindpunt van een klasse noemt men de *klassengrenzen*. Het verschil tussen twee opeenvolgende klassengrenzen wordt de *klassenbreedte* (= b) genoemd.

De totale lengte van het meetinterval is het verschil tussen de hoogste en de laagste meetwaarde: $111 - 81 = 30$ gram. Dit interval verdelen we in een aantal klassen, waarbij de

klassenbreedte in alle gevallen gelijk is. Bij zeer *scheve verdelingen* – bijvoorbeeld een verdeling met veel lage en weinig hoge waarden – worden vaak verschillende klassenbreedten genomen. In ons voorbeeld nemen we respectievelijk een klassenbreedte van 4 en van 10 gram. Dit betekent dat we in het eerste geval 4 meetwaarden samenvoegen, bijvoorbeeld 80 - 83, 84 - 87, 88 - 91, enzovoorts, en in het tweede geval 10 meetwaarden, bijvoorbeeld 80 - 89, 90 - 99, enzovoorts. Vervolgens kan nu elke meetuitkomst ingedeeld worden in een van de voorgeschreven klassen. Als dit is uitgevoerd voor alle meetuitkomsten, kan een histogram worden getekend.

In figuur 3.2 is het histogram getekend voor een klassenbreedte van 4 en in figuur 3.3 voor een klassenbreedte van 10 gram.

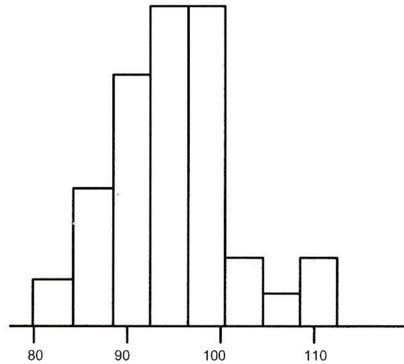


Fig. 3.2 Klassenbreedte $b = 4$

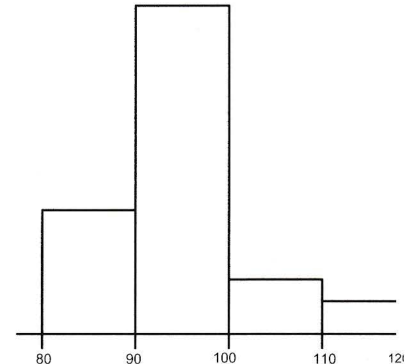


Fig. 3.3 Klassenbreedte $b = 10$

Het histogram in figuur 3.2 geeft een veel beter totaalbeeld dan het histogram in figuur 3.1. Tegenover dit voordeel staat het nadeel dat details verloren zijn gegaan, doordat de oorspronkelijke meetuitkomsten van de steekproef niet meer teruggevonden kunnen worden. In figuur 3.3 is het verlies aan details erg groot, terwijl het histogram in vergelijking met figuur 3.2 er niet duidelijker op geworden is. Dit betekent dat er een optimale indeling in klassen bestaat. De hierbij behorende procedure zullen we nu vastleggen in een voorschrift.

Gegeven is een aantal meetwaarden. Dit aantal geven we aan met de letter n.

1. Bepaal nu eerst het verschil tussen de hoogste en laagste meetwaarde in de steekproef. Dit verschil wordt *spreidingsbreedte* (Eng: *range*) genoemd en wordt genoteerd als R.

2. De klassenbreedte (b) verkrijgen we door eerst de gevonden spreidingsbreedte te delen door de wortel uit het aantal waarnemingen.

$$\frac{\text{range}}{\sqrt{n}} = \frac{R}{\sqrt{n}} \tag{3.1}$$

De uitkomst van deze deling wordt vervolgens volgens bestaande regels afgerond in het aantal decimalen waarin de meetwaarden van de steekproef zijn uitgedrukt. En daarmee hebben we b gevonden.

3. De *klassengrenzen* worden nu gevormd door waarden die *veelvouden* zijn van de gevonden klassenbreedte in punt 2.
4. Ten slotte wordt voor elke klasse een *klassenmidden* bepaald, door het rekenkundig gemiddelde te nemen van de klassenondergrens en klassenbovengrens.

Opmerking

Bij toepassing van bovenstaande ISO-procedure is het aantal klassen ongeveer gelijk aan $\sqrt{n}$. Dit kunnen we dan ook als vuistregel hanteren.

Voor de gegevens uit tabel 3.1 levert dit het volgende resultaat op:

1. $n = 50$ en $R = 111 - 81 = 30$
2. De klassenbreedte is: $b = \dfrac{R}{\sqrt{n}} = \dfrac{30}{\sqrt{50}} = 4{,}24$, afgerond geeft dit $b = 4$ (want de meetuitkomsten zijn in dit geval in eenheden, dus zonder decimalen gegeven).
3. De klassengrenzen worden nu gevormd door getallen die deelbaar zijn door 4. De klassengrenzen van de eerste klasse krijgt men door in de buurt van de laagste meetuitkomst (81) waarden te zoeken, die veelvouden van 4 zijn. De eerste klassenondergrens wordt dan 80 en de bijbehorende klassenbovengrens 84. De laagste meetuitkomst 81 valt in deze klasse. De grenzen van de volgende klassen worden:
$80 + 2 \cdot 4 = 88$
$80 + 3 \cdot 4 = 92$
$80 + 4 \cdot 4 = 96$ enzovoorts.
Nadeel van deze indeling is dat er meetuitkomsten zijn die precies met een klassenbovengrens en de volgende klassenondergrens samenvallen. Zo geeft de waarde 84 moeilijkheden, omdat het niet duidelijk is of deze waarde thuishoort in de klasse 80 - 84 of in de klasse 84 - 88. Deze moeilijkheden hadden we kunnen voorkomen door de meetwaarden niet af te ronden. We moeten ons realiseren dat de meetwaarde 84 in feite een afgerond getal is tussen 83,5 en 84,5. Om nu de vraag te omzeilen in welke klassen de afgeronde waarden 84, 88, 92 enzovoorts thuishoren, worden de klassengrenzen verminderd met de helft van het afrondingsinterval. Let wel: dit is alleen nodig bij discrete variabelen en niet bij continue variabelen. Wanneer we als meting 84,000000 gram hebben gedaan en niet hebben afgerond, kunnen we de klassengrens wel gelijk aan 84 kiezen. Het verschil tussen de waarnemingsuitkomsten 83,999999, 84,000000 en 84,000001 is namelijk te verwaarlozen. In ons voorbeeld zijn de meetuitkomsten weergegeven in eenheden, dus het afrondingsinterval is 1. De klassengrenzen worden nu:
$80 - 0{,}5 = 79{,}5$; $84 - 0{,}5 = 83{,}5$; $88 - 0{,}5 = 87{,}5$ enzovoorts. We krijgen dan de klassen:
$79{,}5 - 83{,}5$ (hier zit 83 dus wel in, 84 niet meer)

83,5 – 87,5
87,5 – 91,5 enzovoorts.

Naast deze notatie komt men vaak een eenvoudiger notatie tegen. Voor de berekende klassenbovengrens zet men het 'kleiner dan' ($<$) -teken. Dit betekent: alle waarnemingen tot aan de betreffende klassenbovengrens.

In ons voorbeeld vinden we dan de volgende klassen:
80– $<$ 84
84– $<$ 88
88– $<$ 92 enzovoorts.

We moeten ons goed realiseren dat we in dit voorbeeld in de klasse 80– $<$ 84 nog steeds alle mogelijke (afgeronde) waarnemingsuitkomsten 80, 81, 82 en 83 kunnen opnemen, en niet de waarnemingsuitkomst 84!

Opmerking

Wanneer de waarnemingsuitkomsten als niet-afgeronde *reële getallen* werden beschouwd, heeft de notatie 80– $<$ 84 een andere betekenis dan in dit geval, waarbij de waarnemingsuitkomsten gehele getallen zijn. Immers, wanneer de meetwaarde een niet-afgerond reëel getal zou zijn geweest, wordt met de klasse 80– $<$ 84 het interval 80,000000 - 84,000000 bedoeld.

4. Het midden van de klassen (*klassenmidden*) krijgen we uit het gemiddelde van de klassenbovengrens en de klassenondergrens, voor de eerste twee klassen zijn dit:

$$\frac{79{,}5 + 83{,}5}{2} = 81{,}5 \quad \text{en} \quad \frac{83{,}5 + 87{,}5}{2} = 82{,}5 \text{ enzovoorts.}$$

Opdracht

Ga na dat de klasse 80– $<$ 84 in dit geval *niet* als klassenmidden $\frac{80 + 84}{2} = 82$ heeft.

Door voor de bovenstaande procedure te kiezen, krijgt men slechts één mogelijke, unieke klassenindeling. Iedereen krijgt, op grond van eenzelfde getallenmassa, ook dezelfde frequentieverdeling, hetgeen voor de praktijk noodzakelijk is. De gevonden (empirische) frequentieverdeling dient als schatting van de populatieverdeling.

Van de frequentietabel 3.2 kunnen we nu de eerste 2 kolommen invullen. Verder moeten we nu alle meetwaarden indelen in de bijbehorende klassen (kolom 3).

Tellen we de aantallen in kolom (3) op, dan vinden we als totaal $n = 50$.

In kolom (4) staan ten slotte de *relatieve frequenties* per klasse. Deze waarden verkrijgt men door het aantal meetwaarden in iedere klasse te delen door het totaal aantal meetwaarden van alle klassen. In ons voorbeeld dus door 50.

Tabel 3.3 Frequentietabel van 50 gewichtsmetingen (klassenbreedte = 4)

(1) klassengrenzen	(2) klassenmidden	(3) frequentie	(4) rel.frequentie
79,5 - 83,5	81,5	1	0,02
83,5 - 87,5	85,5	5	0,10
87,5 - 91,5	89,5	11	0,22
91,5 - 95,5	93,5	14	0,28
95,5 - 99,5	97,5	14	0,28
99,5 - 103,5	101,5	2	0,04
103,5 - 107,5	105,5	1	0,02
107,5 - 111,5	109,5	2	0,04
	totaal	50	1,00

In tabel 3.3 zijn de relatieve frequenties in fracties aangegeven. Met andere woorden, voor elke klasse is het aandeel in het totaal van de waarnemingen opgegeven. Men kan de relatieve frequentie ook in procenten opgeven door de relatieve frequentie met 100 te vermenigvuldigen.
In figuur 3.4 is de frequentieverdeling van tabel 3.3 als histogram getekend.

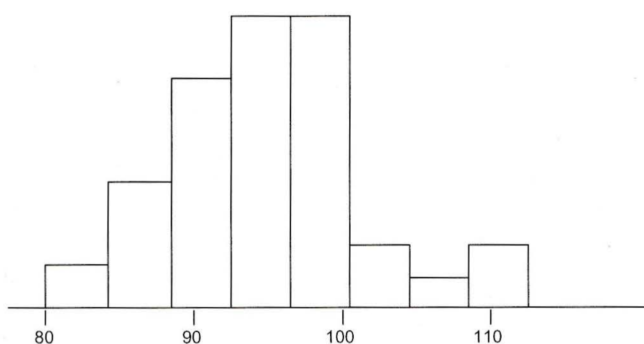


Fig. 3.4 Histogram van 50 gewichtsmetingen (Gegevens tabel 3.3)

Tot nu toe hebben we frequentieverdelingen besproken waarbij alle klassen dezelfde klassenbreedte hebben. Dit gaat op als de uiteindelijke frequentieverdeling redelijk symmetrisch is. Bij erg scheve verdelingen kunnen we de gevolgde procedure niet toepassen en werken we met ongelijke klassenbreedtes. Hierdoor ontstaan zo weinig mogelijk lege klassen. Een voorbeeld hiervan is de inkomstenverdeling. Hoe hoger het inkomen, hoe minder mensen er zijn die dit inkomen verdienen. De inkomstenverdeling is een scheve verdeling. Het heeft in tabel 3.4 geen zin om het betrekkelijk geringe aantal personen met een inkomen van 60.000 euro tot 100.000 euro te verdelen over meerdere klassen.

Tabel 3.4 Verdeling van de belastbare inkomens van personen ($\times$1000 euro)

(1)	(2)	(3)	(4)	(5)
klassen	klassenbreedte	aantal personenen	rel. frequentie	freq. dichtheid
0– < 2	?	663	0,103	?
2– < 4	2	257	0,040	128
4– < 6	2	430	0,067	215
6– < 10	4	1132	0,117	283
10– < 14	4	1121	0,175	280
14– < 20	6	1465	0,228	244
20– < 28	8	763	0,119	95,4
28– < 40	12	333	0,052	27,7
40– < 60	20	152	0,024	7,6
60– < 100	40	63	0,100	1,6
100– < 500	400	30	0,005	0,1
> 500	?	1	0,000	?
	totaal	6410	1,000	

Bij het tekenen van een histogram moet men er in dit geval rekening mee houden dat bij ongelijke klassenbreedten de hoogte van een kolom geen maat is voor het aantal inkomens in de betreffende klasse. Alleen de *oppervlakte* van een kolom is feitelijk een maat voor de frequentie.

De hoogte h van een kolom bij een bepaalde klasse wordt bepaald door het quotiënt van het aantal meetwaarden per klasse en de bijbehorende klassenbreedte. Dit quotiënt wordt de *frequentiedichtheid* genoemd (laatste kolom tabel 3.4). Door bij ongelijke klassenbreedten de frequenties te delen door de verschillende klassenbreedten, wordt de frequentie per klasse gestandaardiseerd naar de eenheid van metingen. De frequentiedichtheid is dus de frequentie per eenheid van de beschouwde grootheid. De klassen kunnen nu beter onderling worden vergeleken. De hoogte van elke kolom komt overeen met de bijbehorende frequentiedichtheid. We zien dit in figuur 3.5 op basis van de gegevens van tabel 3.4.

Hoewel de klasse 14000– < 20000 de klasse is met het hoogste aantal meetwaarden, ligt de top van de verdeling ongeveer bij 10000. De klasse met de hoogste frequentiedichtheid is de klasse 6000– < 10000.

3.2.3 Cumulatieve frequentieverdelingen

Het samenstellen van een frequentietabel uit een groot aantal waarnemingsuitkomsten en het tekenen van het histogram hebben we uitvoerig besproken. Daarbij hebben we uitsluitend gewerkt met de absolute frequentie of met de relatieve frequentie per klasse. In de praktijk werkt men ook vaak met de *cumulatieve frequenties*. Aan de frequentietabel voegen we nog een kolom toe. In deze kolom zetten we het aantal meetwaarden tot en met de

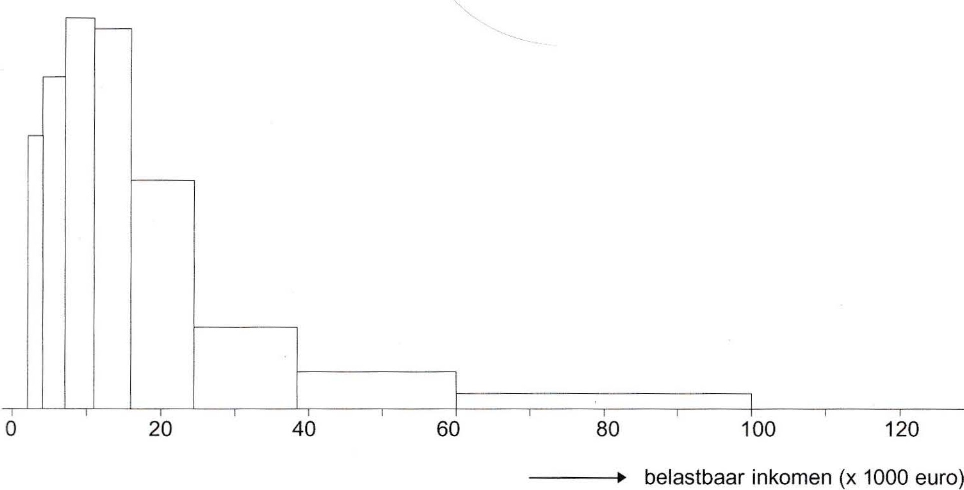


Fig. 3.5 Frequentiedichtheden bij ongelijke klassenbreedte

beschouwde klasse. Per opeenvolgende klasse neemt men dus de som van de frequentie van de betreffende klasse en de totale frequentie van alle voorgaande klassen. We spreken nu van de *gecumuleerde* (gesommeerde) frequenties.
Als voorbeeld nemen we de gegevens van tabel 3.3, waarbij we de frequentietabel uitbreiden met een vijfde kolom voor de cumulatieve frequenties.

Tabel 3.5 Frequentietabel voor de gegevens van voorbeeld 3.1

(1)	(2)	(3)	(4)	(5)
klassen	klassenmidden	frequentie	rel. frequentie	cumulatieve frequentie
79,5 – 83,5	81,5	1	0,02	
83,5 – 87,5	85,5	5	0,10	
87,5 – 91,5	89,5	11	0,22	
91,5 – 95,5	93,5	14	0,28	
95,5 – 99,5	97,5	14	0,28	
99,5 – 103,5	101,5	2	0,04	
103,5 – 107,5	105,5	1	0,02	
107,5 – 111,5	109,5	2	0,04	

In de vijfde kolom zetten we voor de verschillende klassen de cumulatieve frequenties, die als volgt worden verkregen. De cumulatieve frequentie is het aantal meetwaarden in een bepaalde klasse, vermeerderd met het aantal meetwaarden in alle voorgaande klassen.

frequentie	cumulatieve frequenties
1	1
5	5+1=6
11	6+11=17
14	14+17=31
enzovoorts	enzovoorts

Deze gegevens kunnen we nu overbrengen naar kolom 5 van de frequentietabel. Eventueel kunnen we nog een extra kolom toevoegen met de *relatieve cumulatieve* frequenties.

Tabel 3.6 Frequentietabel voor de gegevens van tabel 3.1, inclusief de cumulatieve frequenties

(1)	(2)	(3)	(4)	(5)	(6)
klassengrenzen	klassenmidden	frequentie	rel. freq	cum.freq.	rel. cum.freq.
79,5 – 83,5	81,5	1	0,02	1	0,02
83,5 – 87,5	85,5	5	0,10	6	0,12
87,5 – 91,5	89,5	11	0,22	17	0,34
91,5 – 95,5	93,5	14	0,28	31	0,62
95,5 – 99,5	97,5	14	0,28	45	0,90
99,5 – 103,5	101,5	2	0,04	47	0,94
103,5 – 107,5	105,5	1	0,02	48	0,98
107,5 – 111,5	109,5	2	0,04	50	1,00

Voor de frequentieverdeling in tabel 3.6 geldt bijvoorbeeld dat 17 uitkomsten een waarde hebben kleiner dan de klassenbovengrens 91,5 en dat 31 uitkomsten een lagere waarde hebben dan de klassenbovengrens 95,5, enzovoorts.
Willen we de cumulatieve frequentieverdeling in beeld brengen, dan gaan we daarbij als volgt te werk. De cumulatieve frequenties worden als punten uitgezet boven de betreffende klassen*boven*grens en *niet* boven het klassenmidden. Beneden de waarde 79,5 komen geen uitkomsten voor, dus wordt daar een punt op de horizontale as – de nullijn – geplaatst. Boven de waarde 83,5 zetten we een punt bij 1, bij 91,5 een punt bij 17, enzovoorts. De uitgezette punten worden nu door rechte lijnstukken verbonden.
In de praktijk zetten we echter meestal niet de absolute cumulatieve frequenties uit, maar de relatieve cumulatieve frequentie (meestal in procenten). In figuur 3.6 is dit uitgevoerd voor de gegevens van tabel 3.6.

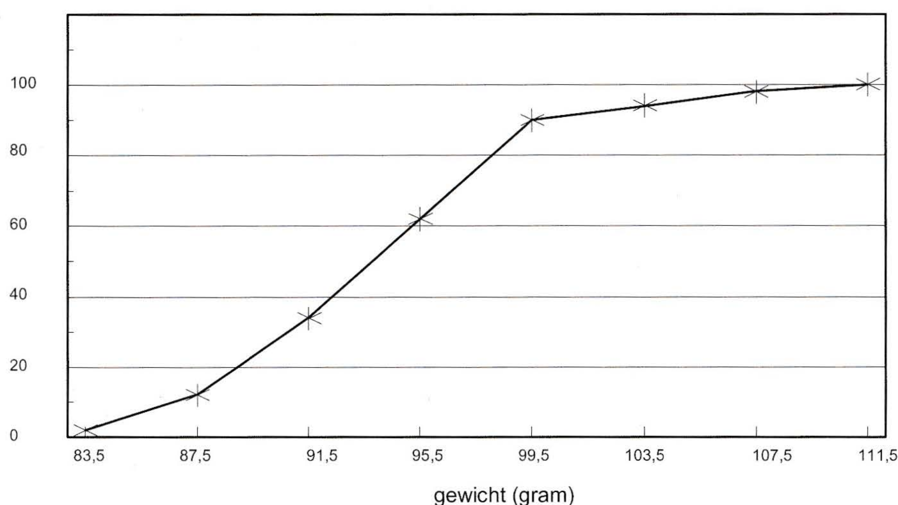


Fig. 3.6 Relatieve cumulatieve frequentieverdeling in procenten

3.2.4 Kwantielen

Uitgaande van de relatieve cumulatieve frequentieverdeling kan men alle meetwaarden verdelen in intervallen met daarin gelijke frequenties. Deze intervallen met gelijke frequenties worden kwantielen genoemd. Daarbij worden de volgende indelingen onderscheiden.

1. Het gehele interval van 0 - 100% op de verticale as verdelen in twee intervallen, ieder met een frequentie van 50%. Het bijbehorende, zogenaamde 50%-punt op de horizontale as wordt de *mediaan* (= Me) genoemd.
2. Het gehele interval op de verticale as indelen in vier intervallen, ieder met een frequentie van 25% van het totaal aantal meetwaarden. Op deze wijze krijgen we op de horizontale as de *kwartielen*. Het eerste kwartiel, aangeduid met Q_1, is het 25%-punt. Onder dit punt ligt 25% van het aantal data. Het tweede kwartiel (Q_2) is het 50%-punt, dit betekent dat 50% van alle data onder het tweede kwartiel ligt. Q_2 is dus gelijk aan de mediaan (Q_2 = Me). Het derde kwartiel (Q_3) is het 75%-punt. Boven het derde kwartiel Q_3 ligt 25% van de verdeling.
3. Wordt het totale interval op de verticale as in tien intervallen verdeeld met elk 10% van de meetwaarden, dan krijgt men op de horizontale as de *decielen*. Onder het eerste deciel (D_1) ligt 10% van de verdeling, onder het tweede deciel (D_2) ligt 20% van de verdeling, enzovoorts.
4. De *percentielen* verdelen het gehele interval op de verticale as in honderd intervallen met elk 1% van de meetwaarden. Het berekenen van percentielen is alleen zinvol bij grote aantallen waarnemingsuitkomsten.

Voorbeeld 2

In figuur 3.7 is de relatieve cumulatieve frequentieverdeling getekend van de lengteverdeling van 60 geproduceerde asjes, met daarin ingetekend de 3 kwartielen.

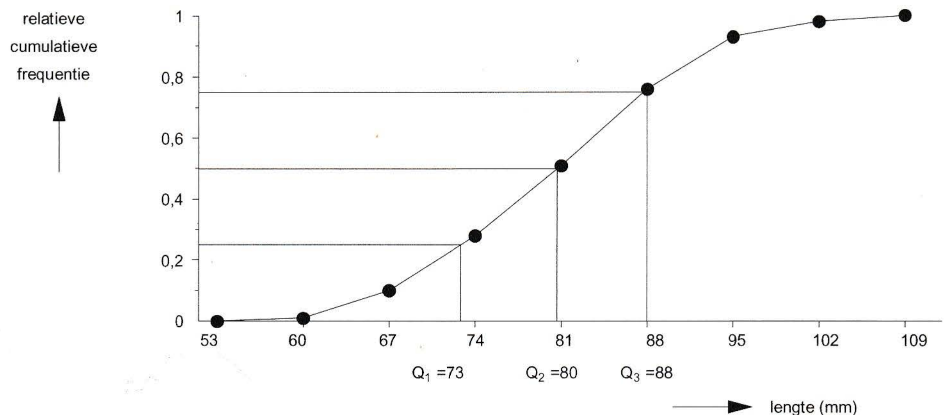


Fig. 3.7 Kwartielen

Het *eerste kwartiel* ligt dus bij het punt 73 mm ($Q_1 = 73$). Dit betekent dat 25% van de lengten een waarde heeft kleiner dan 73 mm.

Men kan nu ook uit de figuur afleiden, dat 50% van de lengten een waarde heeft kleiner dan 80 mm ($Q_2 = \text{Me} = 80$) en 75% van de uitkomsten kleiner is dan 88 mm ($Q_3 = 88$).

3.2.5 Frequentiepolygoon

Een *frequentiepolygoon* ontstaat door de (absolute of relatieve) frequentiedichtheid uit te zetten tegen de *klassenmiddens* en vervolgens deze punten onderling te verbinden door rechte lijnstukken.

Zetten we de absolute waarden uit, dan verkrijgen we de absolute frequentiepolygoon. Zetten we de relatieve frequenties uit, dan spreken we van de relatieve frequentiepolygoon.

Bij de constructie van een frequentiepolygoon is het gebruikelijk om aan beide zijden van het variatiegebied nog een klasse toe te voegen met frequentie nul. De breedte van deze klasse is gelijk aan de breedte van de naastgelegen klasse. Op deze wijze verkrijgen we een frequentiepolygoon die begint en eindigt op de horizontale as (zie de dikkere punten in figuur 3.8). In figuur 3.8 zijn frequentiepolygonen getekend voor zowel een frequentieverdeling met gelijke klassenbreedte als met ongelijke klassenbreedte. Let erop dat in beide gevallen de frequentiedichtheid (en niet de frequentie) op de verticale as tegen de klassenmiddens op de horizontale as is afgezet.

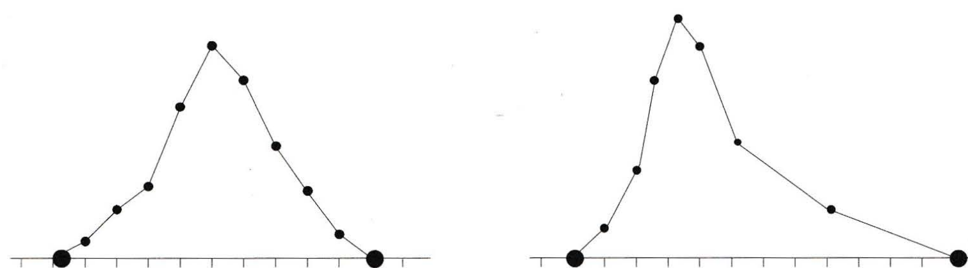


Fig. 3.8 Frequentiepolygoon met gelijke respectievelijk ongelijke klassenbreedten

3.3 Kenmerken voor centrale ligging

In de voorgaande paragrafen hebben we gezien hoe de waarnemingsuitkomsten overzichtelijker gemaakt kunnen worden door het maken van tabellen en grafieken. Voor verdere statistische analyse van de waarnemingsuitkomsten is het nuttig om per variabele de gegevens samen te vatten in frequentietabellen en histogrammen. Maar meestal wil men de uitkomsten per variabele nog beknopter beschrijven en de uitkomsten karakteriseren door één of meer kengetallen (bij steekproeven) of *parameters* (bij populaties). De kentallen van een steekproef zijn, mits de steekproef representatief en aselect is, een goede benadering (schatting) van de parameters van de populatie waaruit de steekproef afkomstig is.
We onderscheiden kengetallen die een maatstaf zijn voor de *ligging* van de waarnemingsuitkomsten en kengetallen die de *mate van spreiding* van de waarnemingsuitkomsten vastleggen. Als eerste groep van kengetallen, bespreken we de kentallen die iets zeggen over de (centrale) ligging van de waarnemingsuitkomsten. We zullen met een voorbeeld beginnen.

Voorbeeld 3

Bij de ingangscontrole van een grondstof voor een productieproces bepaalt men het vastestofgehalte van de grondstof. Daartoe wordt uit de aangevoerde grondstof een steekproef van 10 monsters genomen, waarvan het vastestofgehalte wordt bepaald. Men vindt de volgende waarden:

49,3 49,0 51,0 49,7 50,5 50,1 49,5 50,1 50,7 50,1

Bij het bestuderen van deze 10 uitkomsten, ziet men dat de waarden dicht bij elkaar liggen. De waarden liggen gegroepeerd om een centrale waarde van ongeveer 50.

Centrale waarden of centrumwaarden spelen een belangrijke rol bij het karakteriseren van meetuitkomsten en worden in de praktijk vrij algemeen gebruikt.
In deze paragraaf worden drie centrale waarden (= kentallen/parameters voor de ligging) behandeld, die in de praktijk het meeste worden toegepast, namelijk:

- het *rekenkundig gemiddelde*, meestal kortweg *gemiddelde* genoemd;
- de *mediaan*;
- de *modus*.

De centrale waarden of centrummaten geven een maat voor het midden of centrum van de verdeling van de meetuitkomsten. De bekendste centrummaat is het rekenkundig gemiddelde, maar ook de modus en mediaan worden veel gebruikt. Zij duiden beide aan waar de waarnemingsuitkomsten zich concentreren.

3.3.1 Het rekenkundig gemiddelde

In de statistiek is het (rekenkundig) gemiddelde een belangrijk getal, maar ook in de dagelijkse omgang wordt het regelmatig gebruikt. Voor de schrijfwijze van het rekenkundig gemiddelde maakt men onderscheid tussen de beschrijving van de populatie en het vastleggen van een steekproefresultaat. Het rekenkundig gemiddelde van een populatie wordt genoteerd met de griekse letter μ (spreek uit: *mu*).

Opmerking

De parameters van een populatie worden meestal met een Griekse letter geschreven, dit ter onderscheid van de kentallen van een steekproef.

Zijn de steekproefuitkomsten van n stuks weergegeven door de letters $x_1, x_2, ..., x_n$, dan wordt het rekenkundig gemiddelde van de steekproef (= steekproefgemiddelde) weergegeven door het symbool $\overline{x}$ (spreek uit: 'x-streep').

Definitie

Zijn $x_1, x_2, ..., x_N$ de N waarnemingsuitkomsten van een populatie, dan is het rekenkundig gemiddelde van die populatie:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} \tag{3.2}$$

Definitie

Zijn $x_1, x_2, ..., x_n$ de n waarnemingsuitkomsten van een steekproef, dan is het steekproefgemiddelde:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{3.3}$$

In voorbeeld 3 is het steekproefgemiddelde:

$$\begin{aligned}\overline{x} &= \frac{49{,}3 + 49{,}0 + 51{,}0 + 49{,}7 + 50{,}5 + 50{,}1 + 49{,}5 + 50{,}1 + 50{,}7 + 50{,}1}{10} \\ &= \frac{500}{10} = 50{,}0\end{aligned}$$

Voorbeeld 4

Van 25 personen verdienen er vier 450 euro, twee 950 euro, tien 265 euro, drie 350 euro en zes personen verdienen 311 euro. Wat is het gemiddelde inkomen?

Oplossing

Het is duidelijk dat we in dit geval een snellere berekening krijgen door te rekenen met frequenties en niet door alle 25 uitkomsten op te tellen.

Dus:

$$\begin{aligned} 4 \cdot 450 &= 1800 \\ 2 \cdot 950 &= 1900 \\ 10 \cdot 265 &= 2650 \\ 3 \cdot 350 &= 1050 \\ 6 \cdot 311 &= 1866 \\ \sum_{i=1}^{n} x_i &= 9266 \end{aligned}$$

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{9266}{25} = 370.64 \text{ euro}$$

Het rekenkundig gemiddelde, op de wijze van voorbeeld 4 verkregen, wordt het *gewogen* (rekenkundig) gemiddelde genoemd.
De bedragen worden vermenigvuldigd, of gewogen met de aantallen. De aantallen zijn de weegfactoren.

Definitie

Wanneer de n waarnemingsuitkomsten van een steekproef zijn ondergebracht in frequenties f_k (=aantallen) per uitkomst en er zijn K verschillende uitkomsten x_k ($k = 1, 2, ..., K$), dan geldt voor het steekproefgemiddelde:

$$\overline{x} = \frac{\sum_{k=1}^{K} f_k x_k}{n} \tag{3.4}$$

Het gemiddelde van een frequentieverdeling

Een toepassing van bovenstaande definitie vindt men bij frequentieverdelingen. Eerst worden alle n data van een steekproef op de juiste wijze in klassen verdeeld. Het gemiddelde van de steekproef kan vervolgens *benaderd* worden door voor de weegfactoren in formule (3.4) de frequenties van de klassen (f_k) te nemen. Voor x_k wordt het klassenmidden m_k van de k-de klasse genomen. Dus

$$\overline{x} = \frac{\sum_{k=1}^{K} f_k m_k}{n}$$

Opdracht

Bij de berekening van het gemiddelde van een frequentieverdeling op bovengenoemde wijze wordt ervan uitgegaan dat de klassenmiddens het gemiddelde zijn van de waarnemingsuitkomsten in de betreffende klasse. Op welke veronderstelling is deze aanname gebaseerd?

Het rekenkundig gemiddelde van een steekproef is, mits de steekproef representatief is, een goede schatting van het rekenkundig gemiddelde van de populatie waaruit die steekproef afkomstig is. Hoe goed deze schatting is wordt vastgelegd in een zogenaamde betrouwbaarheidsinterval. Hierop komen we in een volgend hoofdstuk (hoofdstuk 8) terug.

3.3.2 De mediaan

Met een andere centrummaat hebben we al kennisgemaakt, namelijk de *mediaan* (= Me). De mediaan verdeelt de waarnemingsreeks in twee intervallen, waarbij 50% van de uitkomsten een waarde heeft die kleiner is dan de mediaan en 50% van de uitkomsten een waarde heeft die groter is dan de mediaan.
De mediaan kan ook als volgt worden gedefinieerd:

Definitie

Als de meetuitkomsten van een waarnemingsreeks gerangschikt worden naar volgorde van grootte, dan is de mediaan bij een *oneven* aantal waarnemingen gelijk aan de *middelste* waarde van de reeks.
Bij een *even* aantal waarnemingen is de mediaan gelijk aan het *rekenkundig gemiddelde* van de twee middelste waarden.

Opmerking

Een reeks van waarnemingsuitkomsten, geschreven in volgorde van meting, noteren we als: $x_1, x_2, x_3, ..., x_n$.
Als deze reeks in volgorde van grootte wordt gerangschikt, dan schrijven we dit als: $x_{(1)}, x_{(2)}, x_{(3)}, ..., x_{(n)}$.
Hierin is $x_{(1)}$ de laagste meetuitkomst en $x_{(n)}$ de hoogste meetuitkomst.

Voorbeeld 5

Een steekproef van 7 meetwaarden geeft de volgende uitkomsten: 3, 4, 6, 9, 1, 4, 9.
De mediaan van deze meetuitkomsten bepalen we door de waarden eerst op volgorde van grootte te sorteren:

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$
1	3	4	4	6	9	9

Hieruit volgt dat de middelste waarnemingsuitkomst Me = 4

Voorbeeld 6

Heeft men de volgende 8 meetuitkomsten: 28, 21, 22, 25, 26, 28, 22, 24.
Na sorteren op volgorde van grootte:

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$
21	22	22	24	25	26	28	28

volgt: $\text{Me} = \dfrac{24 + 25}{2} = 24{,}5$

3.3.3 De modus

Ten slotte hebben we nog een centrale maat: de *modus*.

Definitie

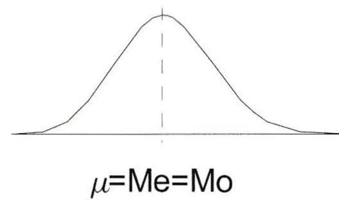
De *modus* (Engels: mode = Mo) is die waarde van de waarnemingsuitkomsten, die het meest voorkomt. Als de waarnemingsuitkomsten zijn ondergebracht in een frequentieverdeling, is de modus het klassenmidden van de klasse met de hoogste frequentiedichtheid. De desbetreffende klasse wordt de *modale klasse* genoemd.

Bij een frequentieverdeling met gelijke klassenbreedtes is de klasse met de hoogste frequentie de modale klasse. Bij frequentieverdelingen met ongelijke klassenbreedtes is de modale klasse de klasse met de hoogste frequentiedichtheid, dus de frequentie van die klasse gedeeld door de klassenbreedte. De modus correspondeert met de *top* van de verdeling.
Het bekendste voorbeeld van de modus als centrale waarde is het ‘modale inkomen’. Het modale inkomen is de salarisklasse met de meeste inkomens.
In tabel 3.3 is het wat moeilijk om een modale klasse aan te wijzen, daar twee klassen dezelfde frequentie hebben, in dit geval kan men als modus de waarde 95,5 euro nemen. Bij tabel 3.4 zijn de klassenbreedten niet gelijk en dan is de klasse met de hoogste frequentiedichtheid de modale klasse. Dus de klasse 6000 - $<$10000 (euro). Het modale inkomen is 8000 euro, namelijk het klassenmidden van deze klasse.

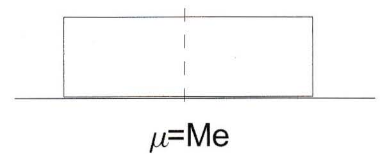
3.3.4 De vergelijking van de verschillende centrumwaarden

Het verband en het verschil tussen het rekenkundig gemiddelde (= μ) de mediaan (= Me) en de modus (= Mo) van een populatie zien we het duidelijkst in de verschillende verdelingsvormen. De vorm van een verdeling wordt meestal voorgesteld door de zogenaamde *ideale kromme*. Hiermee wordt de grafiek bedoeld die de toppen van het frequentiepolygoon op continue wijze met elkaar verbindt.

a. Symmetrische verdelingen:



klokvormige of normale verdeling



rechthoekige of uniforme verdeling

Fig. 3.9a Ideale krommen van symmetrische verdelingen

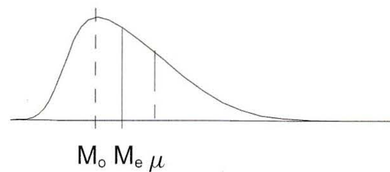
Bij ééntoppige symmetrische verdelingen vallen het rekenkundig gemiddelde, de mediaan en de modus samen. Een normale verdeling heeft de vorm van een (kerk)klok. We zullen deze verdeling zeer vaak tegenkomen.

Opdracht

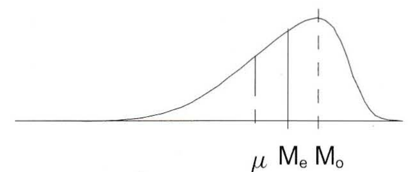
Bedenk drie voorbeelden van een normale verdeling.

Bij een rechthoekige verdeling is er geen sprake van een modus, omdat de mogelijke uitkomsten alle met dezelfde frequentie voorkomen. Voorbeeld van een rechthoekige verdeling is het aantal ogen bij het gooien met een zuivere (dat wil zeggen 'eerlijke') dobbelsteen. Als de dobbelsteen zuiver is, zal elke waarnemingsuitkomst (1, 2, 3, 4, 5, 6) ongeveer even vaak voorkomen.

b. Scheve (asymmetrische) verdelingen:



positief scheve verdeling



negatief scheve verdeling

Fig. 3.9b Ideale kromme van asymmetrische verdelingen

- Voor een *positief scheve* verdeling (staart naar rechts) geldt: Mo $<$ Me $< \mu$
- Voor een *negatief scheve* verdeling (staart naar links) geldt: $\mu <$ Me $<$ Mo

Een voorbeeld van een positief scheve verdeling is de inkomensverdeling in Nederland.

Opdracht

Bedenk een voorbeeld van een negatief scheve verdeling

c. *Meertoppige verdelingen*:

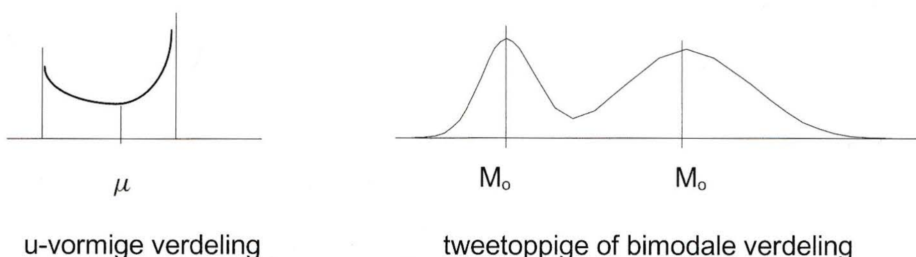


Fig. 3.9c Meertoppige verdelingen

Bij een u-vormige verdeling is het zinloos om één modus weer te geven. Een voorbeeld van een u-vormige verdeling is de bewolkingsgraad (in procenten): helemaal bewolkt en helemaal onbewolkt komen in Nederland vaker voor dan bijvoorbeeld half bewolkt. Bij een tweetoppige (bimodale) of meertoppige verdeling hebben we in principe twee 'modale waarden' en het is zinloos om over een gemiddelde en een mediaan te spreken. Bimodale verdelingen worden gevonden wanneer een populatie uit twee of meer deelpopulaties met verschillende gemiddelden bestaat, of wanneer bijvoorbeeld twee producties van verschillende machines worden samengevoegd.

In het algemeen kan men stellen dat modus en mediaan niet gevoelig zijn voor uitschieters (extreme waarnemingsuitkomsten), dit in tegenstelling tot het rekenkundig gemiddelde.

3.3.5 Verschuiven en vermenigvuldigen

Ten slotte merken we nog het volgende over de in deze paragraaf gedefinieerde centrummaten op: wanneer bij alle waarnemingsuitkomsten van een steekproef (of populatie) hetzelfde getal wordt opgeteld, verschuiven gemiddelde, mediaan en modus over dezelfde afstand naar rechts. Hiermee wordt nog eens duidelijk dat de genoemde kentallen de ligging van de verdeling bepalen. De gehele verdeling verschuift als het ware. Dezelfde eigenschap geldt voor aftrekken. Wanneer van alle data hetzelfde getal wordt afgetrokken, verschuiven gemiddelde, mediaan en modus over dezelfde afstand naar links.
Wanneer alle waarnemingsuitkomsten met hetzelfde getal (> 1) vermenigvuldigd worden, worden gemiddelde, mediaan en modus eveneens met dat getal vermenigvuldigd. Dit betekent dat het histogram van de verdeling wordt opgerekt: de staven worden breder. Een soortgelijke eigenschap geldt ook wanneer alle data worden gedeeld door hetzelfde getal. Wanneer dit getal groter is dan 1, krimpt het histogram: de staven worden smaller.
We kunnen deze eigenschappen in de volgende stelling vastleggen:

Stelling 1

Wanneer $y_i = \dfrac{x_i - A}{B}$ voor alle waarden van i ($i = 1, 2, 3, ...n$), geldt:

1. gemiddelde van $y_i = \dfrac{(\text{gemiddelde van } x_i) - A}{B}$
2. $\text{Me}(y_i) = \dfrac{\text{Me}(x_i) - A}{B}$
3. $\text{Mo}(y_i) = \dfrac{\text{Mo}(x_i) - A}{B}$

Voorbeeld 7

Het gemiddelde van 25 toetscijfers bleek 5,6 te bedragen, de mediaan was 6 en de modus 5.

De docent besloot alle cijfers met een punt op te hogen. Het gemiddelde wordt 6,6, de mediaan 7 en de modus 6.

Wanneer de docent alle cijfers met een factor 1,1 vermenigvuldigd had, was het gemiddelde $1{,}1 \times 5{,}6 = 6{,}16$ geworden. De mediaan zou 6,6 geworden zijn (afgerond een 7) en de modus 5,5 (afgerond een 6).

3.4 Kenmerken van spreiding

In tabel 3.7 zijn de gewichten weergegeven van 10 zakken aardappels, 5 afkomstig van een steekproef uit een partij A en 5 afkomstig van een steekproef uit een partij B.

Tabel 3.7 De gewichten van twee steekproeven uit een partij A en een partij B (in kg)

partij A	partij B
51,3 – 49,0 - 51,0 - 49,7 - 50,5	50,1 - 50,5 - 50,1 - 50,7 - 50,1

De gemiddelden van beide steekproeven zijn gelijk $\overline{x}_A = \overline{x}_B = 50{,}3$ kg, toch verschillen beide steekproeven wezenlijk. De waarnemingsuitkomsten van steekproef A vertonen een grotere spreiding ten opzichte van het gemiddelde dan de uitkomsten van steekproef B. Om een juiste en volledige indruk van beide reeksen te krijgen, is het noodzakelijk om naast een centrummaat, ook een maat voor de spreiding aan te geven.

3.4.1 Spreidingsbreedte

De eenvoudigste maat om een spreiding aan te geven, is de spreidingsbreedte R. Dit kengetal is al eerder aan de orde geweest in paragraaf 3.2.2, bij het opmaken van een frequentieverdeling.

Definitie

De spreidingsbreedte R (Engels: *range*) is het verschil tussen de hoogste en de laagste uitkomst van een reeks waarnemingsuitkomsten.

Voorbeeld 8

De spreidingsbreedten van partij A en van partij B uit tabel 3.7 zijn respectievelijk
$R_A = 51{,}3 - 49{,}0 = 2{,}3$ kg en $R_B = 50{,}7 - 50{,}1 = 0{,}6$ kg
De spreiding van de uitkomsten in de steekproef uit partij A is veel groter, dan de spreiding in de steekproef uit partij B.

Hoewel zeer eenvoudig uit te rekenen, heeft de spreidingsbreedte een belangrijk nadeel: Voor de berekening van de spreidingsbreedte worden alleen de twee uiterste waarden gebruikt. Dit heeft tot gevolg dat niet alle beschikbare informatie wordt benut. De tussenliggende waarden worden niet gebruikt. Bovendien is de spreidingsbreedte zeer gevoelig voor extreem grote of kleine waarden.
De spreidingsmaten waarvoor bovengenoemde nadelen niet gelden, zijn de *variantie* en de *standaardafwijking*.

3.4.2 Variantie

In figuur 3.10 zijn de uitkomsten van de steekproef uit partij A (zie tabel 3.7) uitgezet. We willen nu niet alleen gebruikmaken van de twee uiterste waarden, maar ook van de tussenliggende waarden.

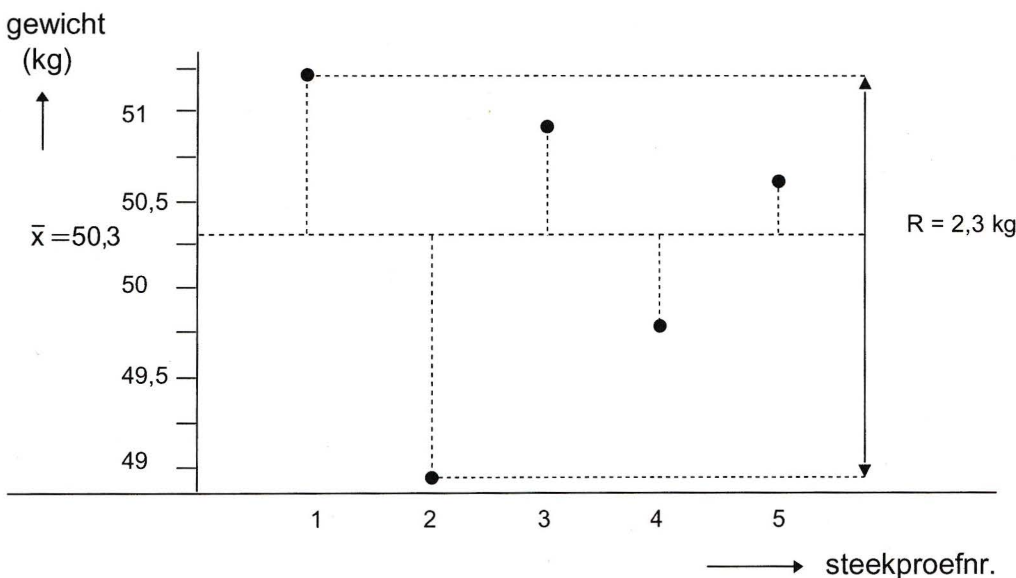


Fig. 3.10 Spreiding gewichten uit partij A

We gaan van elke uitkomst de afwijking ten opzichte van het gemiddelde bepalen $(x_i - \overline{x})$. We krijgen dan:

$x_1 - \overline{x} = 51{,}3 - 50{,}3 = 1{,}0$
$x_2 - \overline{x} = 49{,}0 - 50{,}3 = -1{,}3$
$x_3 - \overline{x} = 51{,}0 - 50{,}3 = 0{,}7$
$x_4 - \overline{x} = 49{,}7 - 50{,}3 = -0{,}6$
$x_5 - \overline{x} = 50{,}5 - 50{,}3 = 0{,}2$

$\sum_{i=1}^{5}(x_i - \bar{x}) = 0.$

Dit is geen toeval want

$$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n\bar{x} = \sum_{i=1}^{n} x_i - n \cdot \frac{\sum_{i=1}^{n} x_i}{n} = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i = 0$$

De som van de afwijkingen ten opzichte van het gemiddelde is dus per definitie gelijk aan nul.

Aan de som van de afwijkingen ten opzichte van $\overline{x}$ hebben we als spreidingsmaat blijkbaar niets. Door de afwijkingen ten opzichte van $\overline{x}$ te kwadrateren en daarna te sommeren, ondervangen we dit probleem. De som van de kwadratische afwijkingen is niet gelijk aan nul (behalve in het uitzonderlijke geval dat alle waarnemingsuitkomsten hetzelfde zijn). Hierdoor is de gemiddelde kwadratische afwijking ten opzichte van $\overline{x}$ te bepalen. En daarmee hebben we de *variantie* gedefinieerd.

Definitie

De variantie is de gemiddelde kwadratische afwijking ten opzichte van het gemiddelde. In formulevorm:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n} \tag{3.5}$$

Notatie: σ^2 = variantie van een populatie

Formule (3.5) geldt slechts in die gevallen waarbij men de afwijkingen ten opzichte van het populatiegemiddelde μ bepaalt. Heeft men echter eerst het steekproefgemiddelde $\overline{x}$ moeten berekenen uit de waarnemingsreeks, dan verandert de berekening in zoverre, dat men niet deelt door n, maar door $n - 1$.

Voor de variantie van een steekproef geldt dus:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} \qquad (3.6)$$

Deze correctie kan men als volgt verklaren. Door eerst $\overline{x}$ te berekenen, zijn niet alle waarden $(x_i - \overline{x})$ onafhankelijk. Er zijn $n - 1$ van de $(x_i - \overline{x})$-waarden vrij te kiezen, maar dan ligt de n-de waarde vast, omdat de som gelijk aan 0 moet zijn. Dit komt omdat door $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ de som van de meetwaarden is vastgelegd.

Men spreekt in dit verband van $n - 1$ *vrijheidsgraden* voor de steekproefvariantie. Het begrip vrijheidsgraad komt men in de statistiek veelvuldig tegen. Op dit moment gaan we er niet verder op in.

Is het populatiegemiddelde μ bekend, dan hoeven we niet eerst $\overline{x}$ te bepalen. De $(x_i - \mu)$-waarden in een aselecte steekproef zijn dan onderling onafhankelijk. Voor de berekening van de variantie heeft men dan ook n vrijheidsgraden.

In de praktijk echter is in de meeste gevallen μ onbekend en zal men de steekproefvariantie moeten bepalen.

Doordat in de formule $s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$ eerst $\bar{x}$ moet worden bepaald, kunnen er afrondingsfouten ontstaan. Via een algebraïsche afleiding is de formule in een beter te hanteren vorm te gieten.

Te bewijzen is dat de teller voor de formule voor de variantie geschreven kan worden als:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

Voor de steekproefvariantie krijgen we dan als tweede formule:

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1} \qquad (3.7)$$

Voorbeeld 9

De uitkomsten van een steekproef zijn 7, 9, 7, 11, 6. De variantie van deze vijf uitkomsten wordt als volgt bepaald:

x_i	x_i^2
7	49
9	81
7	49
11	121
6	36
$\sum_{i=1}^{5} x_i = 40$	$\sum_{i=1}^{n} x_i^2 = 336$

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{336 - \frac{40^2}{5}}{4} = 4{,}00$$

Opmerking

Op een (grafische) zakrekenmachine zijn meestal twee toetsen voor de variantie (of de wortel daaruit, zie hieronder) aanwezig. In het ene geval wordt de formule voor een populatie gebruikt, in het andere geval die van een steekproef. Of zulke toetsen in de praktijk bruikbaar zijn, hangt af van de mogelijkheid te controleren of de juiste meetwaarden zijn ingevoerd. Bij een computerprogramma zoals EXCEL bestaat die mogelijkheid uiteraard wel.

Opmerking

Net als bij de berekening van het gemiddelde kan de variantie van een steekproef geschat worden door eerst een frequentieverdeling te maken. Daarna wordt het gemiddelde geschat op basis van de frequentieverdeling (zie de opmerking na formule (3.4)). Vervolgens worden in formule (3.6) of formule (3.7) alle waarnemingsuitkomsten x_i van een bepaalde klasse 'samengevat' door het klassenmidden m_k van die klasse, terwijl als weegfactor de frequentie van die klasse gebruikt wordt. Formule (3.6) gaat op deze wijze over in:

$$s^2 = \frac{\sum_{k=1}^{K} f_k (m_k - \overline{x})^2}{n-1}$$

waarbij K het aantal klassen is, m_k het klassenmidden van de k-de klasse en f_k de frequentie van die klasse.

3.4.3 De standaardafwijking

Het nadeel van de variantie als spreidingsmaat is dat de variantie gedefinieerd is als de som van het kwadraat van de afwijkingen, waardoor ook de eenheid, waarin de variantie wordt

uitgedrukt, kwadratisch is. Deze eenheid is moeilijk te interpreteren. Vandaar dat we vaak de wortel uit de variantie als spreidingskental nemen.

Definitie
De *standaardafwijking* (Engels: *standarddeviation*) is de wortel uit de variantie, dus voor de steekproefvariantie geldt:

$$s = \sqrt{s^2} \tag{3.8}$$

De berekening van de standaardafwijking gaat via de berekening van de variantie, waarna de wortel van de variantie wordt genomen. Voor een populatie geldt uiteraard hetzelfde, alleen is de notatie anders: σ.

3.4.4 De variatiecoëfficiënt

De tot nu toe besproken spreidingsmaten zijn zogenaamde absolute spreidingsmaten. Het nadeel van de variantie en de standaardafwijking als spreidingsmaat is dat ze gevoelig zijn voor de dimensie van de meetuitkomsten. Als bijvoorbeeld van meters op centimeters wordt overgegaan, wordt de standaardafwijking 100 keer zo groot en de variantie zelfs 10.000 keer. Daarom wordt in de praktijk vaak nog een andere spreidingsmaat gebruikt, namelijk de variatiecoëfficiënt (= γ voor een populatie en c voor een steekproef).

Definitie
De variatiecoëfficiënt (γ voor een populatie en c voor een steekproef) is het quotiënt van de standaardafwijking en het gemiddelde:

$$\gamma = \frac{\sigma}{\mu} \text{ respectievelijk } c = \frac{s}{\bar{x}} \tag{3.9}$$

De variatiecoëfficiënt is een dimensieloze spreidingsmaat. Dit betekent dat de grootte van de variatiecoëfficiënt niet afhangt van de dimensie van de meeteenheid. Meestal wordt de variatiecoëfficiënt dan ook in procenten weergegeven: $\gamma = \frac{\sigma}{\mu} \times 100\%$ of $c = \frac{s}{\bar{x}} \times 100\%$
Uit de gegeven definitie blijkt dat de standaardafwijking van een reeks waarnemingsuitkomsten, door middel van de variatiecoëfficiënt, uitgedrukt wordt in een fractie (of in procenten) van het rekenkundig gemiddelde.

3.4.5 Verschuiven en vermenigvuldigen (2)

Wanneer bij alle waarnemingsuitkomsten van een populatie of steekproef hetzelfde getal wordt opgeteld, heeft dit geen invloed op de spreidingsmaatstaven variantie, standaardafwijking en range, maar wel op de variatiecoëfficiënt. Verschuiven van de verdeling (ook naar links, bij aftrekken) heeft alleen invloed op de ligging, niet op de mate van spreiding.
Wanneer de waarnemingsuitkomsten alle met hetzelfde getal vermenigvuldigd (of door dat getal gedeeld: delen is het omgekeerde van vermenigvuldigen!) worden, heeft dit wel in-

vloed op de spreidingsmaatstaven. Op de variantie is het effect sterker dan op de standaardafwijking. De variatiecoëfficiënt verandert niet (waarom niet?). We leggen deze eigenschap vast in de volgende stelling:

Stelling 2
Wanneer $y_i = \dfrac{x_i - A}{B}$ voor alle waarden van i $(i = 1, 2, 3, ...n)$, geldt:

1. range van $y_i = \dfrac{(\text{range van } x_i)}{B}$
2. $\text{standaardafwijking}(y_i) = \dfrac{\text{standaardafwijking}(x_i)}{B}$
3. $\text{variantie}(y_i) = \dfrac{\text{variantie}(x_i)}{B^2}$

3.5 De verwachtingswaarde

Bij kansverdelingen, die we later zullen bespreken, gebruikt men voor het rekenkundig gemiddelde van de populatie (of verdeling) vaak de uitdrukking *verwachtingswaarde* of kortweg *verwachting* van de verdeling. De verwachtingswaarde of verwachting (Engels: *expectation*) hangt samen met de *experimentele wet van de grote aantallen.* Als men oneindig vaak een experiment herhaalt, nadert de uitkomst van zo'n experiment naar een constante waarde. Zo zal bij het oneindig vaak opgooien van een munt, men gemiddeld genomen in 50% van de gevallen 'kop' boven krijgen en in 50% van de gevallen 'munt'. We zeggen nu dat de verwachtingswaarde van 'het aantal keren kop' bij het werpen met een munt 0.5 of 50% is. Dit wordt genoteerd als: E('kop') = 0.5 (de letter E is afgeleid van *expectation).* Zo is de verwachtingswaarde van het aantal keren dat zes wordt gegooid bij het werpen met een dobbelsteen: E('zes') = $\frac{1}{6}$.
In zijn algemeenheid kan men zeggen:
Wanneer een experiment n keer wordt uitgevoerd en X het gemeten kenmerk (bijvoorbeeld de lengte) is en x_i $(i = 1, ..., n)$ zijn de bij X behorende uitkomsten van n experimenten, geldt voor de verwachtingswaarde (= rekenkundig gemiddelde) van X:

$$E(X) = \frac{\sum_{i=1}^{n} x_i}{n} = \mu \tag{3.10}$$

In analogie hiermee kunnen we voor X^2 (met waarde x_i^2, voor $i = 1, ..., n$) schrijven:

$$E(X^2) = \frac{\sum_{i=1}^{n} x_i^2}{n} \tag{3.11}$$

Duiden we de variantie van kenmerk X aan met $var(X)$, dan kan men afleiden:

$$var(X) = E(X^2) - \{E(X)\}^2 \tag{3.12}$$

De formules voor de verwachtingswaarde (dan wel het gemiddelde) en de variantie van een verdeling zullen we later in dit boek nog verschillende keren tegenkomen.

Opgaven

1. Als gemiddelde van 5 metingen vond men $\bar{x} = 23{,}0$. Toen men later de spreiding wilde berekenen, waren er slechts 4 van de 5 meetwaarden terug te vinden:
15,0 - 27,0 - 19,0 - 35,0 - ?
a. Bereken de ontbrekende waarde.
b. Bereken vervolgens de standaardafwijking.

2. Van een vulmachine voor poedervormige producten worden van de lopende band aselect 10 pakjes gewogen. De volgende gewichten zijn gevonden:
52,3 - 53,6 - 51,5 - 53,8 - 51,2 - 50,9 - 55,0 - 52,4 - 52,3 - 55,9
a. Bereken de mediaan en het gemiddelde.
b. Bereken de spreidingsbreedte en de standaardafwijking.
c. Bereken de variatiecoëfficiënt.

3. Op 12 achtereenvolgende dagen wordt de temperatuur van het koelwater van een chemisch proces gemeten. De gevonden waarden zijn:
33 - 24 - 39 - 48 - 26 - 35 - 38 - 54 - 23 - 34 - 29 en 37
a. Bereken de gemiddelde temperatuur en de spreiding (standaardafwijking);
b. Bereken de variatiecoëfficiënt.

4. Van een proces wordt de temperatuur gemeten in graden Celsius (°C). De gegevens over een maand gezien gaven het volgende beeld:

$$\bar{x} = 54 \text{ en } s_x = 2{,}35$$

Men wil de temperatuur uitdrukken in graden Fahrenheit (°F). Bereken nu de gemiddelde temperatuur, de standaardafwijking en de variatiecoëfficiënt in graden Fahrenheit.
$(°F = \frac{9}{5}°C + 32)$.

5. Van een verzameling waarnemingsuitkomsten is het gemiddelde gelijk aan $\bar{x} = 25$ en de standaardafwijking gelijk aan $s = 2{,}4$.

Bereken het gemiddelde en de standaardafwijking van de verzameling wanneer men iedere waarnemingsuitkomst:

a. met 2,5 vermindert.
b. door 3 deelt.
c. met 2,5 vermeerdert.
d. met 3 vermenigvuldigt.
e. eerst met 2,5 vermindert en daarna door 3 deelt.
f. eerst door 3 deelt en daarna met 2,5 vermindert.
g. eerst met 2,5 vermeerdert en daarna met 3 vermenigvuldigt.
h. eerst met 3 vermenigvuldigt en daarna met 2,5 vermeerdert.

6. In onderstaande tabel zijn de gegevens opgenomen van de dagelijkse zwaveloxyde-uitstoot van een bepaalde energiecentrale. De gegevens hebben betrekking op een periode van drie maanden (SO_2 in tonnen).

15,8	26,4	17,3	11,2	23,9	24,8	18,7	13,9	9,0	13,2
22,7	9,8	6,0	14,7	17,5	26,1	12,8	28,6	17,6	23,7
26,8	22,7	18,0	20,5	11,0	20,9	15,5	19,4	16,7	10,7
19,1	15,2	22,9	26,6	20,4	21,4	19,2	21,6	16,9	19,0
18,5	23,0	24,6	20,1	16,2	18,0	7,7	13,5	23,5	14,5
14,4	29,6	19,4	17,0	20,8	24,3	22,5	24,6	18,4	18,1
8,3	21,9	12,3	22,3	13,3	11,8	19,3	20,0	25,7	32,8
25,9	10,5	15,9	27,5	18,1	17,9	9,4	24,1	20,1	28,5

Stel van deze gegevens een frequentietabel op en teken het bijbehorende histogram. Bereken vervolgens het gemiddelde en de standaardafwijking uit de frequentietabel.

7. Bij een radarcontrole door de gemeentepolitie te Alkmaar werden op een bepaald punt binnen de bebouwde kom in een uur tijd achtereenvolgens de volgende snelheden (in km/u) gemeten:

51	56	48	71	65	80	39	45	58	67
105	56	62	45	68	56	55	67	70	85
45	43	52	57	68	60	61	75	44	57
34	45	56	54	48	60	50	51	76	45
42	51	44	44	42	55	60	70	59	50

a. Bepaal de modus en bereken het gemiddelde en de mediaan van de ongegroepeerde waarnemingsuitkomsten.
b. Bepaal hun range.
c. Stel een frequentietabel samen en teken het bijbehorende histogram.
d. Bepaal de modus en bereken het gemiddelde en de mediaan van de frequentieverdeling.

e. Geef commentaar op de drie hetzij sub a hetzij sub d berekende maatstaven voor ligging.

8. Metalen pennen bestemd voor de montage in buizen, worden geëtst in een etsvloeistof. Bij het onderdompelen in de etsvloeistof worden de pennen ingeklemd in een houder. Na het etsen wordt het blanke (niet geëtste) uiteinde gemeten (lengte in mm). De meetresultaten zijn als volgt:

0,60	0,67	0,78	0,71	0,64	0,71	0,74	0,71	0,62	0,70
0,72	0,68	0,64	0,71	0,69	0,63	0,71	0,70	0,70	0,68
0,70	0,70	0,66	0,67	0,70	0,70	0,74	0,69	0,70	0,72
0,66	0,72	0,70	0,71	0,73	0,75	0,69	0,72	0,75	0,71
0,75	0,65	0,66	0,70	0,80	0,68	0,70	0,67	0,70	0,66

Stel van deze gegevens een frequentietabel samen en teken het bijbehorende histogram. Bereken vervolgens uit de frequentietabel het gemiddelde en de standaardafwijking door de klassenmiddens als representanten te nemen van alle waarnemingsuitkomsten in de betreffende klasse en de frequentie van de bijbehorende klasse als weegfactoren te nemen in de formules voor gemiddelde en standaardafwijking.

9. In de volgende tabel zijn de gegevens opgenomen van het aantal medewerkers van een bedrijf dat op een bepaalde dag afwezig is. De gegevens berusten op waarnemingen over een periode van 50 dagen.

13	5	13	27	10	16	2	11	6	12
8	21	12	11	7	7	9	16	29	18
3	11	19	6	15	10	14	10	7	24
11	3	6	10	4	6	28	9	12	7
28	12	9	19	8	20	15	5	17	10

a. Stel een frequentietabel op, ook de kolom met de relatieve frequenties, alsook de cumulatieve frequenties.

b. Teken de frequentiepolygoon van de relatieve cumulatieve frequenties.

c. Bereken uit de frequentietabel het gemiddelde aantal afwezigen per dag. Bepaal eveneens de standaardafwijking van het aantal afwezigen per dag.

d. Bepaal de mediaan van het aantal afwezigen per dag en de modus.

4 Kansrekening

4.1 Inleiding

In de vorige hoofdstukken hebben we kennisgemaakt met een aantal begrippen uit de beschrijvende statistiek, namelijk het verzamelen, het rangschikken en het karakteriseren van data. Wanneer we ons gaan bezighouden met de verdere analyse van deze data met de bedoeling daaruit verantwoorde conclusies te kunnen trekken, begeven we ons op het terrein van de toegepaste statistiek. De toegepaste statistiek is gebaseerd op de kansrekening die dan gezien kan worden als de verbindende schakel tussen de beschrijvende statistiek en de toegepaste statistiek. We zullen daarom in dit hoofdstuk aandacht aan deze kansrekening besteden. Daarbij zullen we ons beperken tot de belangrijkste regels die nodig zijn om conclusies, die in de komende hoofdstukken getrokken zullen worden, te kunnen begrijpen. Maar eerst moeten we het begrip 'kans' definiëren. Voor het begrip 'kans' bestaan verschillende definities.

4.2 De verschillende definities van het begrip kans

Men zou kunnen zeggen dat er verschillende invalshoeken zijn om het begrip 'kans' te definiëren. We zullen dit illustreren met enkele voorbeelden:

Voorbeeld 1
Er worden twee zuivere (= 'eerlijke') dobbelstenen tegelijk geworpen. Wat is de kans dat in totaal 7 ogen gegooid worden?

Oplossing
In dit voorbeeld kan de *klassieke definitie* van het begrip 'kans' worden toegepast. We kunnen de mogelijke uitkomsten van het experiment met hun kans van optreden voorspellen, omdat we het gedrag van een dobbelsteen kennen. Een dobbelsteen heeft zes

zijvlakken, die alle zes met even grote kans boven komen. Bij een worp met twee dobbelstenen tegelijk zijn er dus 6 x 6 = 36 *mogelijke* totaaluitkomsten (zie de tabel) met elk dezelfde kans van optreden. In onderstaande tabel staat van alle mogelijke uitkomsten van de worp met twee dobbelstenen hun som (s).

	1	2	3	4	5	6
1	$s = 2$	$s = 3$	$s = 4$	$s = 5$	$s = 6$	$s = 7$
2	$s = 3$	$s = 4$	$s = 5$	$s = 6$	$s = 7$	$s = 8$
3	$s = 4$	$s = 5$	$s = 6$	$s = 7$	$s = 8$	$s = 9$
4	$s = 5$	$s = 6$	$s = 7$	$s = 8$	$s = 9$	$s = 10$
5	$s = 6$	$s = 7$	$s = 8$	$s = 9$	$s = 10$	$s = 11$
6	$s = 7$	$s = 8$	$s = 9$	$s = 10$	$s = 11$	$s = 12$

Het totale aantal ogen is 7 bij de combinaties (3, 4), (4, 3), (5, 2), (2, 5), (1, 6) en (6, 1). Er zijn dus 6 van de 36 combinaties die voldoen aan het kenmerk: 'het totaal aantal ogen is 7'. Deze 6 combinaties noemt men 'gunstig' voor het optreden van de gebeurtenis met dat kenmerk. De kans dat het ogentotaal 7 is, is gelijk aan het aantal gunstige uitkomsten gedeeld door het aantal mogelijke uitkomsten, oftewel $\frac{6}{36} = \frac{1}{6}$. Deze kans kunnen we dus uitrekenen zonder het experiment te hoeven uitvoeren.

4.2.1 De klassieke kansdefinitie

De klassieke definitie die we in bovenstaand voorbeeld hebben toegepast, luidt in algemene vorm:

Definitie

Stel een experiment kent M mogelijke, even waarschijnlijke uitkomsten. De kans op het optreden van een bepaalde gebeurtenis (X) is gelijk aan het aantal uitkomsten $G(X)$ (G van 'gunstig') waarbij die gebeurtenis optreedt, gedeeld door het aantal mogelijke uitkomsten (M):

$$P(X) = \frac{G(X)}{M} \tag{4.1}$$

Populair kunnen we de klassieke kansdefinitie (toegeschreven aan de wiskundige Pierre Simon Laplace; 1749-1827) als volgt formuleren:

$$\text{kans} = \frac{\text{aantal gunstige gevallen}}{\text{aantal mogelijke gevallen}} \tag{4.2}$$

De klassieke definitie van het begrip kans is zeer eenvoudig te hanteren. En waar mogelijk zullen we deze definitie ook gebruiken.

4.2.2 Kans als relatieve frequentie

Een tweede manier om het begrip kans te definiëren, leiden we in met het volgende voorbeeld.

Voorbeeld 2

Er wordt 1000 keer met een dobbelsteen geworpen. Er verschijnt 200 keer een ogenaantal van 6. Wat is de kans dat met die dobbelsteen in een willekeurige worp 6 ogen gegooid worden?

Oplossing

In dit voorbeeld is een experiment, het opwerpen van een dobbelsteen, 1000 maal uitgevoerd. Op het eerste gezicht een zinloos tijdverdrijf, maar bij nader inzien is dit experiment zo vaak uitgevoerd om vast te kunnen stellen of de dobbelsteen zuiver (eerlijk) of onzuiver (oneerlijk) is. Het aantal keren dat het experiment '6' als ogenaantal opleverde, is 200. We kunnen nu de kans dat de geworpen dobbelsteen in één willekeurige worp '6' als uitkomst oplevert, *schatten* door het aantal malen dat het experiment '6' opleverde te delen door het aantal malen dat het experiment werd uitgevoerd: $\frac{200}{1000} = \frac{1}{5}$. De hier toegepaste definitie noemen we de *relatieve frequentie-definitie*, omdat de relatieve frequentie van het aantal keren '6' bepaald werd. De vraag of de dobbelsteen waarmee geworpen is zuiver is of niet, kunnen we nog niet beantwoorden. Het zou kunnen zijn dat bij 10.000 maal gooien de relatieve frequentie van het aantal malen '6' veel dichter bij $\frac{1}{6}$ ligt (dit is volgens de klassieke definitie de kans dat de dobbelsteen bij één worp het ogenaantal '6' oplevert, althans indien de dobbelsteen echt zuiver is, dus alle uitkomsten even grote kans van optreden hebben). Maar het is ook mogelijk dat de gebruikte dobbelsteen wel degelijk onzuiver is. In hoofdstuk 9 zullen we op dit probleem terugkomen.

4.2.3 De wet van de grote aantallen

Het in het vorige voorbeeld beschreven verschijnsel is gebaseerd op de zogenaamde *experimentele wet van de grote aantallen*. Deze wet is als volgt te formuleren:

Stelling 1

Neem aan dat de omstandigheden bij een reeks experimenten niet veranderen. Dan zal de relatieve frequentie waarmee gebeurtenis A optreedt, bij voortdurende toename van het aantal experimenten naderen tot een constante waarde.

Als de dobbelsteen uit de experimenten van voorbeeld 2 inderdaad zuiver is, mogen we verwachten dat de relatieve frequentie van het aantal keren dat 6 ogen gegooid wordt, op den duur naar $\frac{1}{6}$ zal naderen. Zie figuur 4.1 voor een toelichting.
De algemene formulering van de in voorbeeld 2 toegepaste relatieve frequentie-definitie is als volgt:

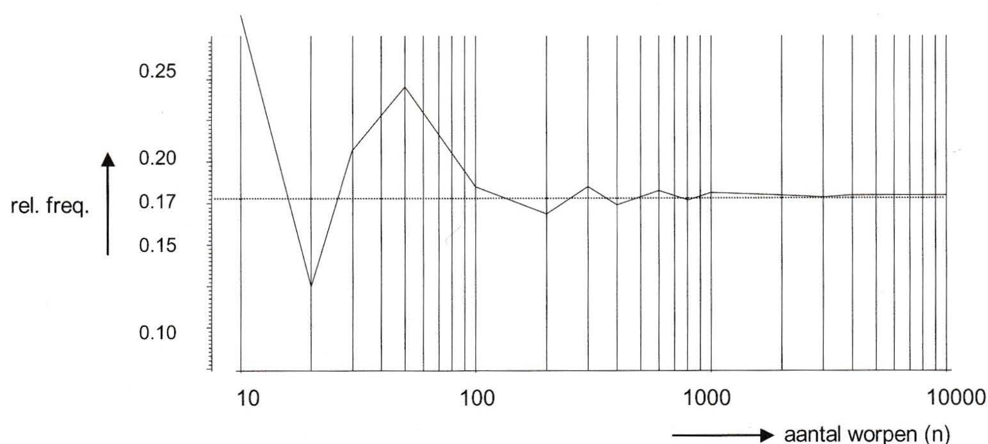


Fig. 4.1 Experimentele wet van de grote aantallen

Definitie

Wanneer een experiment onder gelijke omstandigheden N maal wordt uitgevoerd, is de kans op het optreden van een bepaalde gebeurtenis (X) gelijk aan het aantal malen $n(X)$ dat deze gebeurtenis optreedt, gedeeld door N:

$$P(X) = \frac{n(X)}{N} \tag{4.3}$$

Vaak kennen we de mogelijke uitkomsten van een experiment zonder het te hoeven uitvoeren. Wanneer we tevens voldoende kennis hebben over de kans van optreden van de mogelijke uitkomsten van dat experiment, is het zinloos dat experiment daadwerkelijk uit te voeren en gebruiken we de klassieke kans-definitie. Als we van tevoren hadden geweten dat de dobbelsteen uit voorbeeld 2 zuiver was, hadden we ons de moeite van het werpen kunnen besparen.

4.2.4 Subjectieve kansdefinitie

Vaak is het onmogelijk op volkomen objectieve wijze (met een formele definitie) een kans te definiëren. Toch wordt het kansbegrip dan wel degelijk toegepast, maar we moeten er wat voorzichtiger mee omgaan.

Voorbeeld 3

Een nieuw product wordt ontworpen. Wat is de kans dat dit product een succes wordt?

Oplossing

In dit voorbeeld is het onmogelijk op objectieve wijze de kans uit te rekenen. Zonder de behoefte aan een nieuw product te kennen, is de kans dat het bewuste product een succes wordt dan ook subjectief. We spreken over een subjectieve kans. Wanneer de behoefte aan het product met behulp van marketingtechnieken min of meer bekend is, wordt het iets gemakkelijker om zo'n kans te schatten. Er zal echter vrijwel altijd een gebrek aan voldoende kennis van de 'markt' zijn om zo'n kans volkomen objectief te kunnen bepalen.

In voorbeeld 3 kunnen we nog wel van een experiment spreken, namelijk het ontwerpen van een nieuw product, dat wel of niet goed verkocht zal worden. De mogelijke uitkomsten van dat experiment zijn 'wel een succes' of 'geen succes'. Maar over de kans van optreden van de beide uitkomsten kunnen we zonder een marketingonderzoek weinig zeggen. Bovendien kunnen we zo'n experiment niet herhalen onder dezelfde omstandigheden. Om te kunnen vaststellen of een kans op objectieve wijze kan worden berekend, definiëren we het begrip kansexperiment.

Kansexperiment

In de voorbeelden 1 en 2 was sprake van een experiment, dat bij herhaling onder dezelfde omstandigheden kon worden uitgevoerd. De mogelijke uitkomsten waren bekend en omdat de omstandigheden iedere keer dezelfde waren, verandert de voorspelbaarheid van de verschillende gebeurtenissen niet, ook al voert men het experiment gedurende lange tijd uit. Zo'n experiment heet een *kansexperiment*. Wanneer we op objectieve wijze over kansen willen spreken, moet sprake zijn van een kansexperiment.

We besluiten deze inleiding met nog twee voorbeelden waarin een van de genoemde kansdefinities kan worden toegepast.

Voorbeeld 4

De ervaring heeft geleerd dat van de productie van een bepaald product gemiddeld 5% van het aantal stuks kwalitatief slecht is. Per dag worden zo'n 1000 stuks van dat product geproduceerd. Men neemt een steekproef van 20 stuks. Wat is de kans dat er precies één kwalitatief slecht product bij zit?

Oplossing

In dit voorbeeld is ook weer sprake van een experiment dat iedere dag onder dezelfde omstandigheden verricht zou kunnen worden. De mogelijke uitkomsten van dat experiment zijn 0, 1, 2, ..., 19 of alle 20 kwalitatief slechte producten, òf, wat op hetzelfde neerkomt: 20, 19, 18, ..., 1 of 0 kwalitatief goede producten. De kans op 0, 1, 2,..., 19 of 20 slechte producten kan berekend worden zonder het experiment te hoeven uitvoeren, dus volgens de klassieke definitie. Maar deze kansen zijn niet zo gemakkelijk als in de voorbeelden 1 en 2 te berekenen. Daarvoor is kennis nodig van de rekenregels die in de

kansrekening gebruikt mogen worden. In de volgende paragrafen zullen we deze rekenregels formuleren. In het volgende hoofdstuk (bij de binomiale kansverdeling) zullen we de in dit voorbeeld gevraagde kans daadwerkelijk leren uitrekenen.
Let op: Denk niet dat er bij een steekproef van 20 producten altijd wel één is van kwalitatief slechte aard (5% van 20). De gevraagde kans zou dan 1 zijn ('altijd'). Deze redenering is onjuist: daarbij wordt immers uitgesloten dat er 0, 2 , 3,....19, of 20 producten van slechte kwaliteit zijn, terwijl deze uitkomsten van de steekproef wel degelijk tot de mogelijkheden behoren.

Voorbeeld 5

Een randomgenerator genereert willekeurige reële getallen tussen -1 en 1. Men genereert met zo'n randomgenerator n punten met de coördinaten (x, y), die liggen binnen het vierkant dat gevormd wordt door de punten $(-1, -1)$, $(1, -1)$, $(1, 1)$ en $(-1, 1)$ in het xy-vlak. Dit vierkant heeft dus als zijde $z = 2$. Men telt het aantal punten p dat binnen of op de rand van de cirkel met de vergelijking $x^2 + y^2 = 1$ valt (zie figuur 4.2). Men kan op deze manier de verhouding van de oppervlakken van de cirkel ($= \pi r^2 = \pi$, want $r = 1$) en het vierkant benaderen en daarmee een schatting maken van het getal π:

$$\frac{p}{n} \approx \frac{\pi r^2}{z^2} = \tfrac{1}{4}\pi, \text{ dus } \pi \approx \frac{4p}{n}$$

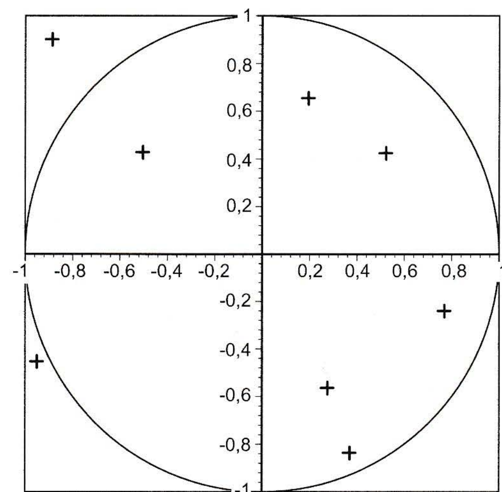


Fig. 4.2 Monte Carlo-simulatie

In dit voorbeeld werd de relatieve frequentie-definitie gebruikt. Hoe groter het aantal punten dat gegenereerd wordt, hoe beter de benadering van het bijzondere getal π. We zien hier een voorbeeld van zogenaamde *Monte Carlo-simulatie*.

4.3 Rekenen met kansen

Met kansen kan tot op zekere hoogte gerekend worden. Hiervoor staat een aantal regels ter beschikking. Deze regels kunnen we aanschouwelijk maken, maar om werkelijk met kansen te kunnen manipuleren, moeten we deze regels netjes formuleren. Daarbij zullen we de uit de wiskunde bekende verzamelingenleer gebruiken. We hebben daartoe te maken met de volgende begrippen en notaties.

4.3.1 De begrippen uitkomstenruimte en gebeurtenis

De *uitkomstenruimte* van een experiment is de verzameling van alle mogelijke uitkomsten van dat experiment. De uitkomstenruimte wordt vaak weergegeven met de hoofdletter U. Zo wordt de uitkomstenruimte van een worp met twee dobbelstenen (voorbeeld 1, uit de inleidende paragraaf) weergegeven als de verzameling: $U_1 = \{(1,1), (1,2), (1,3), ..., (6,6)\}$. Deze verzameling bevat 36 mogelijke uitkomsten.
De uitkomstenruimte van het experiment uit voorbeeld 2 (worp met een dobbelsteen) is de verzameling $U_2 = \{1, 2, 3, 4, 5 \text{ en } 6\}$. In voorbeeld 3 (uitbrengen van een nieuw product) is de uitkomstenruimte van het experiment de verzameling U_3 ={wel een succes, geen succes} en de uitkomstenruimte in voorbeeld 4 (het aantal kwalitatief slechte exemplaren in een steekproef van 20 stuks) is de verzameling $U_4 = \{0, 1, 2, ..., 20\}$. In het vijfde voorbeeld (het genereren van een punt binnen een vierkant) bestaat de uitkomstenverzameling uit alle (oneindig veel) punten in het beschreven vierkant.
Bij de berekening van een kans zijn we geïnteresseerd in die uitkomsten die aan de omschreven kenmerken voldoen. Die uitkomsten vormen altijd een deelverzameling van de uitkomstenruimte. De deelverzameling van uitkomsten die hoort bij de beschrijving, wordt meestal een *gebeurtenis* genoemd.

Voorbeeld 6

Zo kunnen we bij de worp met een dobbelsteen (met uitkomstenruimte $U = \{1, 2, 3, 4, 5, 6\}$) als mogelijke gebeurtenissen formuleren: $V_1 = \{1\}$, $V_2 = \{1 \text{ of } 5\}$, V_3 = {even ogenaantal}, V_4 = {ogenaantal groter dan 3} enzovoorts.

4.3.2 Venn-diagram

Een uitkomstenruimte met verschillende gebeurtenissen kunnen we weergeven in een zogenaamd *Venn-diagram*. In figuur 4.3 is een Venn-diagram getekend voor de uitkomstenruimte behorend bij voorbeeld 6 (de worp met een dobbelsteen), met daarin de gebeurtenissen V_1, V_2, V_3 en V_4.

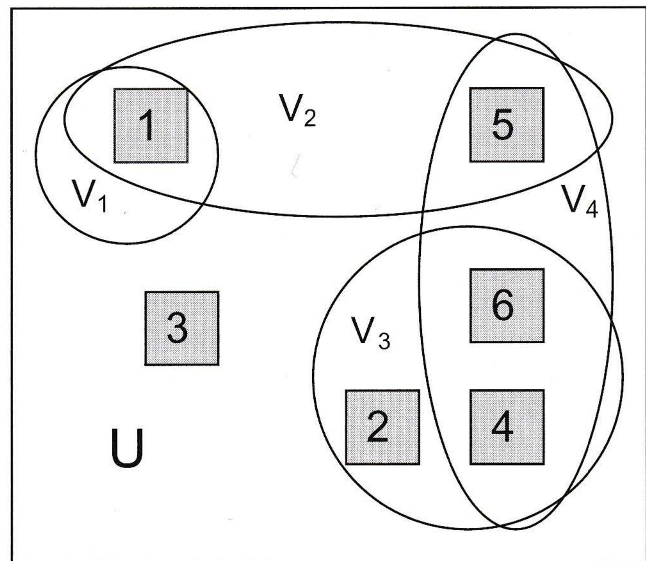


Fig. 4.3 Venn-diagram

4.3.3 Begrippen uit de verzamelingsleer

Uit de verzamelingstheorie worden de volgende notaties gebruikt. Wanneer A en B gebeurtenissen zijn, die deel uitmaken van een uitkomstenruimte U, noteren we

- $A \subset B$, dat wil zeggen A is een *deelverzameling* van B.
- $A \cup B$, de *vereniging* van A en B, dit betreft de uitkomsten die bij gebeurtenis A en/of gebeurtenis B optreden.
- $A \cap B$, de *doorsnede* van A en B, dit zijn de uitkomsten die zowel bij gebeurtenis A als B optreden.
- $\overline{A}$, het *complement* van A, dit is de verzameling van alle uitkomsten die niet bij gebeurtenis A optreden.

Ter toelichting van de zojuist genoemde begrippen uit de verzamelingsleer geven we het volgende voorbeeld.

Voorbeeld 7

$U = \{$alle 52 kaarten in een volledig kaartspel$\}$
$A = \{$de 13 harten in het spel$\}$
$B = \{$de 4 azen in het spel)$\}$
Dan is:

$A \cup B$ = {de 13 harten (inclusief hartenaas) + de 3 overige azen}= $B \cup A$ = {de 4 azen (inclusief hartenaas) + de 12 overige harten}
$A \cap B$ = {hartenaas}
$\overline{A}$ = {de 13 schoppen + de 13 klaveren + de 13 ruiten}
Ter illustratie geven we het bijbehorende Venn-diagram:

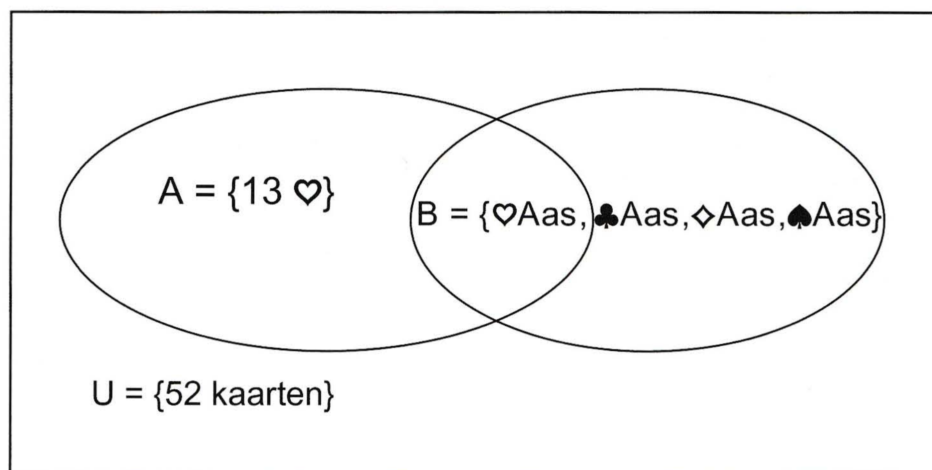


Fig. 4.4 Venn-diagram bij voorbeeld 7

4.3.4 $n \times m$-tabellen

In de kansrekening worden verzamelingen en deelverzamelingen vaak gedefinieerd vanuit een tabel.

Voorbeeld 8

In onderstaande tabel is een groep van 850 studenten verdeeld in vier deelgroepen, afhankelijk van geslacht en gekozen studierichting.

Tabel 4.1 Aantallen vrouwelijke en mannelijke eerstejaarsstudenten in de studierichtingen Bouwkunde en Elektrotechniek aan een Technische Universiteit

	bouwkunde (B)	elektrotechniek (E)	totaal
vrouwelijk (V)	200	50	250
mannelijk (M)	150	450	600
totaal	350	500	850

We kunnen vanuit de tabel definiëren:
U = {alle (850) eerstejaarsstudenten bouwkunde en elektrotechniek};
B = {alle (350) eerstejaarsstudenten bouwkunde};
E = {alle (500) eerstejaarsstudenten elektrotechniek};
V = {alle (250) vrouwelijke eerstejaarsstudenten};
M = {alle (350) mannelijke eerstejaarsstudenten};
Dan is bijvoorbeeld
$B \cap V$ ={alle (200) vrouwelijke eerstejaarsstudenten bouwkunde};
$M \cup B$ = {alle (600) mannelijke eerstejaarsstudenten + alle (200) vrouwelijke eerstejaars bouwkundestudenten }
$\overline{B}$ ={alle (500) eerstejaarsstudenten die geen bouwkunde studeren} = {alle 500 eerstejaarsstudenten die elektrotechniek studeren} = E
$\overline{B \cap V}$ = {de (650) eerstejaarsstudenten die niet tegelijk bouwkunde studeren en vrouw zijn} = {de (250) eerstejaars elektro-studenten (inclusief de vrouwelijke) + alle (400) mannelijke eerstejaars bouwkunde-studenten}

In tabel 4.2 is een aantal deelverzamelingen, afgeleid van tabel 4.1, symbolisch afgebeeld.

Tabel 4.2

	bouwkunde	elektrotechniek	totaal
vrouwelijk	$V \cap B$	$V \cap E$	V
mannelijk	$M \cap B$	$M \cap E$	M
totaal	B	E	U

Opdracht

Definieer zelf in woorden met behulp van de bovenstaande tabellen de verzamelingen $B \cup V$, $\overline{B \cup M}$, $\overline{B} \cup \overline{M}$ en $B \cap E$. Bereken het aantal studenten in de betreffende verzamelingen.

Aantallen elementen in een verzameling

Wanneer we het aantal elementen van een verzameling X aanduiden met $n(X)$, blijkt uit tabel 4.1:
$n(U) = 850$, $n(B) = 350$, $n(E) = 500$, $n(M) = 600$, $n(V) = 250$ en verder:
$n(B \cap V) = 200$, $n(B \cap M) = 150$, $n(E \cap V) = 50$ en $n(E \cap M) = 450$.
Bedenk dat bijvoorbeeld $n(B) = n(B \cap V) + n(B \cap M) = 200 + 150 = 350$.
Willen we vaststellen hoeveel bijvoorbeeld $n(B \cup V)$ bedraagt, dan kunnen we in overeenstemming met de eerder gegeven definitie van $B \cup V$ schrijven:

$n(B \cup V) = n(B) + n(E \cap V) = 350 + 50 = 400.$
Omdat (zie tabel 4.2) $n(E \cap V) = n(V) - n(B \cap V)$
kan $n(B \cup V)$ ook geschreven worden als:
$n(B \cup V) = n(B) + n(V) - n(B \cap V) = 350 + 250 - 200 = 400.$
En ten slotte: omdat $B \cup V$ en $E \cap M$ elkaars complement zijn (bouwkunde en/of vrouw is *niet* electrotechniek en *niet* man), geldt ook nog
dat $n(B \cup V) = n(U) - n(E \cap M) = 850 - 450 = 400.$
Nu we dit soort formules (eventueel met behulp van Venn-diagrammen) kunnen opschrijven, is het een kleine stap terug naar de kansrekening. Op de zojuist ontwikkelde formules komen we straks terug.

4.4 Het formele kansbegrip

Nu we gebeurtenissen kunnen identificeren met verzamelingen en de bijbehorende uitkomsten met elementen uit die verzamelingen, kunnen we het kansbegrip op een formele manier herdefiniëren. We zullen ervoor zorgen dat deze definitie aansluit bij de in de inleiding van dit hoofdstuk gegeven definities van het begrip kans. Het formele kansbegrip omvat drie uitgangspunten (in de wiskunde noemt men dit *axioma*'s) die ons in staat stellen kansen te berekenen met gebruikmaking van wiskundige logica, gebaseerd op verzamelingenleer. We zullen deze drie uitgangspunten hier in het kort bespreken.

Axioma 1
Wanneer bij een kansexperiment de n mogelijke gebeurtenissen K_1, K_2, K_3, ...K_n kunnen optreden, kan aan iedere gebeurtenis K_i een kans $\mathrm{P}(K_i)$ worden toegekend, zodanig dat

$$0 \leq P(K_i) \leq 1 \qquad (4.4)$$

Opmerking
Hier wordt dus gesteld dat een kans nooit negatief kan zijn en ook niet groter dan 1.

Opdracht
Ga na dat deze axiomatische definitie van het begrip kans in overeenstemming is met de twee definities in paragraaf 4.2.

4.4.1 Elkaar uitsluitende gebeurtenissen

Elkaar uitsluitende (ook wel genoemd: *disjuncte*) gebeurtenissen zijn gebeurtenissen die bij eenmalige uitvoering van een kansexperiment niet tegelijkertijd kunnen optreden.
Voorbeelden van elkaar uitsluitende gebeurtenissen zijn:

- 'kop' en 'munt' bij de worp met een muntstuk;
- 'één slecht product' en 'twee slechte producten' bij een kwaliteitscontrole;
- '5 ogen' en 'minder dan 3 ogen' bij de worp met een dobbelsteen;
- 'klaver' en 'schoppen' bij het trekken van een speelkaart.

Elkaar niet-uitsluitende gebeurtenissen zijn gebeurtenissen die bij eenmalige uitvoering van een kansexperiment wél tegelijkertijd kunnen optreden.
Voorbeelden van elkaar niet uitsluitende gebeurtenissen zijn:

- '2 ogen' en 'een even aantal ogen' bij de worp met een dobbelsteen;
- 'één slecht produkt' en 'minder dan twee slechte produkten' bij een kwaliteitscontrole;
- 'ruiten' en 'boer' bij het trekken van een speelkaart.

Axioma 2
Wanneer alle mogelijke gebeurtenissen $K_1, K_2, K_3, ...K_n$ van een kansexperiment elkaars optreden uitsluiten, geldt

$$\sum_{i=1}^{n} P(K_i) = P(K_1) + P(K_2) + ... + P(K_n) = 1 \tag{4.5}$$

Dit axioma is een logisch gevolg van de definitie van het begrip uitkomstenruimte. Wanneer $K_1, K_2, K_3, ...K_n$ alle (elkaar uitsluitende) gebeurtenissen uit de uitkomstenruimte zijn, is de som van de kansen hierop gelijk aan 1. Zie het volgende voorbeeld.

Voorbeeld 9
Als uitkomstenruimte van een worp met een dobbelsteen kunnen we definiëren: {1, 2, 3, 4, 5, 6}. De 6 bijbehorende uitkomsten sluiten elkaar uit en omdat ieder van hen kans $P(K_i) = \frac{1}{6}$ $(i = 1, 2, 3, .., 6)$ heeft, is de som van hun kansen gelijk aan 1.
Pas op! De definitie van de uitkomstenruimte bij een gebeurtenis is meestal niet uniek, dat wil zeggen: er zijn andere uitkomstenruimten mogelijk. Wanneer we als uitkomstenruimte van een worp met een dobbelsteen definiëren: $\{K_1, K_2\}$ met K_1 = {minder dan 5 ogen} en K_2 = {5 ogen of meer}, is eveneens aan axioma 2 voldaan. Ga dit na!

4.4.2 De speciale optelregel

Met een derde axioma wordt de formele definitie van het kansbegrip voltooid.

Voorbeeld 10
Een kansexperiment bestaat uit het gooien met een zuivere dobbelsteen.
Met K_1 = {een even aantal ogen} en K_2 = {5 ogen} geldt er:

$P(K_1 \cup K_2) = P(2,4,6) + P(5) = \frac{3}{6} + \frac{1}{6} = \frac{4}{6} = \frac{2}{3}$.
Maar: met K_1 = {een oneven aantal ogen} en K_2 = {5 ogen} geldt er:
$P(K_1 \cup K_2) = P(1,3,5) = \frac{3}{6}$ en niet $P(K_1 \cup K_2) = P(1,3,5) + P(5) = \frac{3}{6} + \frac{1}{6}$, omdat we $P(5)$ in dat geval dubbel zouden tellen. We mogen de afzonderlijke kansen $P(K_1)$ en $P(K_2)$ blijkbaar alleen optellen wanneer K_1 en K_2 elkaars optreden uitsluiten.

Dit leidt tot de volgende veronderstelling:

Axioma 3
Wanneer de gebeurtenissen K_1 en K_2 elkaar uitsluiten, is de kans dat K_1 of K_2 optreedt gelijk aan de som van hun kansen. In formulevorm

$$P(K_1 \cup K_2) = P(K_1) + P(K_2) \qquad (4.6)$$

In feite definieert $K_1 \cup K_2$ de verzameling van alle elementen die óf tot K_1 óf tot K_2 óf tot beide behoren. Maar omdat $K_1 \cap K_2$ in dit geval leeg is (ze sluiten elkaar immers uit), definieert $K_1 \cup K_2$ hier de verzameling van alle elementen, die ofwel tot K_1 ofwel tot K_2 behoren. We kunnen dit als een speciaal geval beschouwen en spreken daarom ook van de 'speciale' optelregel. Het 'algemene' geval, waarin K_1 en K_2 elkaar niet uitsluiten – dus het geval waarin $K_1 \cap K_2$ niet leeg is – maakt de berekening van $P(K_1 \cup K_2)$ iets gecompliceerder. We komen hierop terug in paragraaf 4.5.

4.5 Rekenregels

Uit de 3 axioma's die we in het kader van het formele kansbegrip in paragraaf 4.3 hebben behandeld, volgt een aantal (nieuwe) rekenregels. Met deze regels is het mogelijk ingewikkelder kansvraagstukken op te lossen dan tot nu toe aan de orde zijn geweest. We zullen die rekenregels in deze paragraaf behandelen.

4.5.1 De complementregel

De eerste regel die volgt uit de axioma's uit de vorige paragraaf is de *complementregel.*

Stelling 2
Voor iedere gebeurtenis K geldt:

$$P(\overline{K}) = 1 - P(K) \qquad (4.7)$$

Deze regel volgt rechtstreeks uit axioma 2 want K en $\overline{K}$ vormen tezamen de uitkomstenruimte en sluiten elkaar uit, zodat $P(K) + P(\overline{K}) = 1$.
Het verdient aanbeveling de complementregel toe te passen wanneer men $P(K)$ moet berekenen in een situatie waarin het minder werk blijkt te zijn om eerst $P(\overline{K})$ te berekenen.

Voorbeeld 11

Bereken de kans dat de som van het aantal ogen in 2 opeenvolgende worpen met een zuivere dobbelsteen minstens 3 bedraagt.

Oplossing

Met i = 'het aantal ogen in de eerste worp' en j = 'het aantal ogen in de tweede worp' is het complement van de gebeurtenis $K = \{i + j \geq 3\}$ identiek aan de gebeurtenis $\overline{K} = \{i + j \leq 2\}$. De gebeurtenis $\overline{K}$ kan alleen gerealiseerd worden met de uitkomst waarbij $i = 1$ en $j = 1$, zodat $n(\overline{K}) = 1$.

Omdat de uitkomstenruimte uit $6 \times 6 = 36$ uitkomsten bestaat, is $n(U) = 36$ en dus geldt er

$$P(K) = 1 - P(\overline{K}) = 1 - \frac{n(\overline{K})}{n(U)} = 1 - \tfrac{1}{36} = \tfrac{35}{36}.$$

4.5.2 De algemene optelregel

Een uitbreiding van axioma 3 is de zogenaamde *algemene optelregel*.

Stelling 3

Wanneer K_1 en K_2 willekeurige gebeurtenissen zijn, geldt de algemene optelregel:

$$P(K_1 \cup K_2) = P(K_1) + P(K_2) - P(K_1 \cap K_2) \tag{4.8}$$

Wanneer $K_1 \cap K_2$ leeg is (dus geen elementen bevat), is $P(K_1 \cap K_2) = 0$ en gaat de algemene optelregel over in de speciale optelregel van axioma 3.

Het principe van de algemene optelregel is het beste te illustreren met een Venn-diagram.

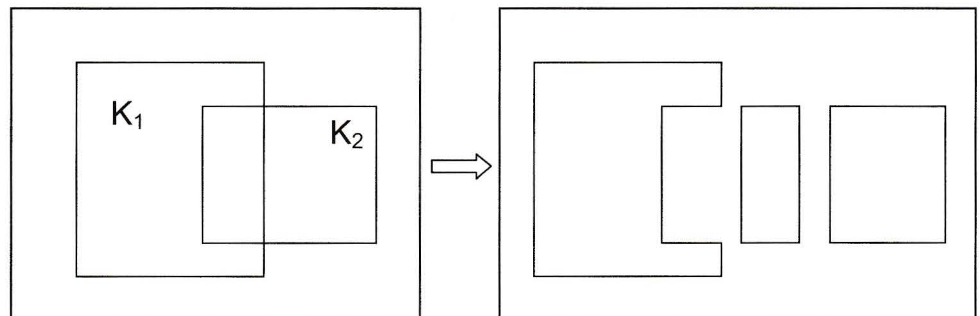


Fig. 4.5 Toelichting principe algemene optelregel

In het linkergedeelte van figuur 4.5 zien we de vereniging $K_1 \cup K_2$ van de verzamelingen K_1 en K_2. In het rechtergedeelte is de vereniging opgesplitst in de verzamelingen $K_1 \cap \overline{K}_2$ (links), $K_1 \cap K_2$ (midden) en $\overline{K}_1 \cap K_2$ (rechts).

Er geldt dus $P(K_1 \cup K_2) = P(K_1 \cap \overline{K_2}) + P(K_1 \cap K_2) + P(\overline{K_1} \cap K_2)$. Verder geldt dat K_1 is op te splitsen in de verzamelingen $K_1 \cap \overline{K_2}$ en $K_1 \cap K_2$. We kunnen dus schrijven $P(K_1) = P(K_1 \cap \overline{K_2}) + P(K_1 \cap K_2)$, oftewel $P(K_1 \cap \overline{K_2}) = P(K_1) - P(K_1 \cap K_2)$.
Op dezelfde manier blijkt ook dat $P(K_2) = P(\overline{K_1} \cap K_2) + P(K_1 \cap K_2)$, dus $P(\overline{K_1} \cap K_2) = P(K_2) - P(K_1 \cap K_2)$.
Conclusie:

$$\begin{aligned} P(K_1 \cup K_2) &= P(K_1) - P(K_1 \cap K_2) + P(K_1 \cap K_2) + P(K_2) - P(K_1 \cap K_2) \\ &= P(K_1) + P(K_2) - P(K_1 \cap K_2) \end{aligned}$$

We zullen regel (4.8) illustreren met een voorbeeld.

Voorbeeld 12
Uit een goed geschud volledig kaartspel (52 kaarten) trekt men een kaart. Hoe groot is de kans op een harten of een boer of beide?

Oplossing
Met K_1 = {harten} en K_2 = {boer},
$K_1 \cup K_2$ = {alle harten (inclusief hartenboer) + de 3 overige boeren} en
$K_1 \cap K_2$ = {hartenboer},
vinden we met behulp van rekenregel (4.8):
P (harten of boer of beide) $= P(K_1 \cup K_2) = P(K_1) + P(K_2) - P(K_1 \cap K_2) =$
P (harten) + P (boer) $-P$ (hartenboer) $= \frac{13}{52} + \frac{4}{52} - \frac{1}{52} = \frac{16}{52} = \frac{4}{13}$.

4.5.3 Voorwaardelijke kansen

Het begrip voorwaardelijke kans speelt in de kansrekening een belangrijke rol: vele kansproblemen zijn slechts op te lossen met gebruikmaking van voorwaardelijke kansen. We zullen eerst het begrip via twee voorbeelden introduceren en toelichten en daarna een rekenregel voor voorwaardelijke kansen formuleren.

Voorbeeld 13
We gebruiken de gegevens van voorbeeld 8. Een groep van 850 studenten is op twee manieren verdeeld: over twee studierichtingen en en naar geslacht. In de tabel staan de aantallen.

	bouwkunde (B)	elektrotechniek (E)	totaal
vrouwelijk (V)	200	50	250
mannelijk (M)	150	450	600
totaal	350	500	850

Bereken de volgende kansen:

a. De kans dat een willekeurig gekozen student uit de groep van 850 studenten een vrouw is.

b. De kans dat een willekeurig gekozen bouwkundestudent vrouw is.

c. De kans dat een willekeurig gekozen student bouwkunde studeert èn vrouw is.

d. De kans dat een willekeurig gekozen vrouw bouwkunde studeert.

Oplossing

We berekenen de volgende kansen:

a. De kans dat een willekeurig gekozen student uit de groep van 850 studenten een vrouw is.

Volgens de klassieke definitie van het kansbegrip is deze kans $P(V) = \dfrac{n(V)}{N} = \frac{250}{850} = \frac{5}{17}$

b. De kans dat een willekeurig gekozen bouwkundestudent vrouw is.

Dit is een zogenaamde voorwaardelijke kans. Gegeven (voorwaarde!) is dat de student bouwkunde studeert. De uitkomstenruimte is dus beperkt tot de groep bouwkundestudenten. In de formule voor de klassieke kansdefinitie is N in de noemer nu niet 850 maar 350. Notatie: $P(V \mid B) = \dfrac{n(B \cap V)}{n(B)} = \frac{200}{350} = \frac{4}{7}$.

c. De kans dat een willekeurig gekozen student bouwkunde studeert én vrouw is.

Ook deze kans wordt met de klassieke definitie berekend. Nu is de uitkomstenruimte wel de verzameling van 850 studenten.

$P(B \cap V) = \dfrac{n(B \cap V)}{N} = \frac{200}{850} = \frac{4}{17}$.

d. De kans dat een willekeurig gekozen vrouw bouwkunde studeert.

Ook hierbij hebben we te maken met een voorwaardelijke kans. De uitkomstenruimte met alle mogelijke gebeurtenissen is nu de verzameling vrouwen:

$P(B \mid V) = \dfrac{n(B \cap V)}{n(V)} = \frac{200}{250} = \frac{4}{5}$.

Merk op dat $P(V \mid B)$ niet gelijk is aan $P(B \mid V)$!!

Merk ook op dat $P(V) \cdot P(B \mid V) = P(B \cap V)$ (zie ook later bij regel (4.9) t/m (4.11)).

Uit het laatste voorbeeld halen we de definitie:

Definitie

De voorwaardelijke kans dat gebeurtenis K_1 optreedt op voorwaarde dat gebeurtenis K_2 optreedt (of is opgetreden) schrijven we als $P(K_1 \mid K_2)$. Spreek uit als: $P(K_1$ onder voorwaarde $K_2)$.

In het volgende voorbeeld worden bijna alle tot dusver ontwikkelde kansregels toegepast.

Voorbeeld 14

Van een partij producten is 90% van goede kwaliteit (G), maar 10% vertoont gebreken ($\overline{G}$). Beschouw dit als een praktijkgegeven. Alle producten worden gekeurd, maar de keuringsdeskundige maakt fouten (zoals ieder mens!). Uit ervaring is bekend dat van de goede producten 5% ten onrechte wordt afgekeurd (dus 95% van de goede producten wordt terecht goedgekeurd). Tevens is bekend dat van de kwalitatief slechte producten ten onrechte toch nog 12% wordt goedgekeurd (dus 88% van de slechte producten wordt terecht afgekeurd). Bereken de kans dat een afgekeurd product toch goed blijkt te zijn.

Oplossing

Gegeven:

$P(G) = 0,9 \Rightarrow P(\overline{G}) = 0,1$

$P(\text{afkeur}|\ G) = 0{,}05 \Rightarrow P(\text{goedkeur}|\ G) = 0{,}95$

$P(\text{goedkeur}|\ \overline{G}) = 0{,}12 \Rightarrow P(\text{afkeur}|\ \overline{G}) = 0{,}88.$

Merk op dat de complementregel drie keer is gebruikt. Merk ook op dat in tegenstelling tot het vorige voorbeeld geen aantallen gegeven zijn, maar kansen (eventueel om te zetten naar percentages). De gevraagde kans berekenen we door te doen alsof de aantallen wel bekend zijn. Vervolgens manipuleren we met de aantallen zodanig dat het kansen worden. Vergelijk de hierbij gebruikte formules met die uit het vorige voorbeeld.

$$\begin{aligned}
P(G \mid \text{afkeur}) &= \frac{\text{het aantal goede (èn) afgekeurde producten}}{\text{het aantal afgekeurde producten}} \\
&= \frac{n(G\ \cap\ \text{afkeur})}{n(\ \text{afkeur})} \\
&= \frac{\dfrac{n(G\ \cap\ \text{afkeur})}{n(\text{totaal})}}{\dfrac{n(\ \text{afkeur})}{n(\text{totaal})}} \\
&= \frac{P(G\ \cap\ \text{afkeur})}{P(\text{afkeur})}
\end{aligned}$$

Merk weer op dat $P(G \cap \text{afkeur})$ blijkbaar te schrijven is als $P(\text{afkeur}) \cdot P(G\ |\text{afkeur})$.

Op soortgelijke wijze kunnen we aantonen dat

$$P(\text{afkeur} \mid G) = \frac{P(G \cap \text{afkeur})}{P(G)}$$

dus

$$P(G \cap \text{afkeur}) = P(G) \cdot P(\text{afkeur} \mid G)$$

en evenzo is

$$P(\text{afkeur} \cap \overline{G}) = P(\overline{G}) \cdot P(\text{afkeur} \mid \overline{G})$$

Bedenken we dat de afgekeurde producten uit twee categorieeën bestaan, namelijk de goede afgekeurde producten en de slechte afgekeurde producten, dan kunnen we volgens de speciale somregel schrijven:

$$\begin{aligned} P(\text{afkeur}) &= P(\text{afkeur} \cap G) + P(\text{afkeur} \cap \overline{G}) \\ &= P(G) \cdot P(\text{afkeur} \mid G) + P(\overline{G}) \cdot P(\text{afkeur} \mid \overline{G}) \end{aligned}$$

Invullen van de gegevens levert

$$P(G \mid \text{afkeur}) = \frac{(0,9)(0,05)}{(0,9)(0,05) + (0,1)(0,88)} = \frac{45}{133} \approx 0,34$$

We kunnen de berekende resultaten ook nu in een (2×2)−tabel aangeven. Alleen staan er in de tabel nu geen aantallen, maar het principe is hetzelfde:

	goedkeuring	afkeuring	totaal
G	$P(G \cap \text{goedkeur})$	$P(G \cap \text{afkeur})$	0,9
$\overline{G}$	$P(\overline{G} \cap \text{goedkeur})$	$P(\overline{G} \cap \text{afkeur})$	0,1
totaal			1

en ingevuld met behulp van de gegevens:

	goedkeuring	afkeuring	totaal
G	(0,9)(0,95) = 0,855	(0,9)(0,05) = 0,045	0,9
$\overline{G}$	(0,1)(0,12) = 0,012	(0,1)(0,88) = 0,088	0,1
totaal	0,867	0,133	1

We hebben in bovenstaande twee voorbeelden kunnen zien hoe een voorwaardelijke kans kan worden berekend. Dit vatten we samen in de volgende rekenregel:

Stelling 4

De kans dat gebeurtenis K_2 zal optreden onder de voorwaarde dat gebeurtenis K_1 optreedt is gelijk aan:

$$P(K_2 \mid K_1) = \frac{P(K_1 \cap K_2)}{P(K_1)} \tag{4.9}$$

Op analoge wijze geldt:

$$P(K_1 \mid K_2) = \frac{P(K_1 \cap K_2)}{P(K_2)}$$

4.5.4 De algemene productregel

Denkend aan hetgeen we in de laatste twee voorbeelden behandeld hebben en de rekenregel (4.9), is het nu niet moeilijk meer de zogenaamde *algemene productregel* te formuleren:

Stelling 5

Wanneer K_1 en K_2 gebeurtenissen zijn met $P(K_1) \neq 0$ en $P(K_2) \neq 0$ dan is:

$$\begin{aligned} P(K_1 \cap K_2) &= P(K_1) \cdot P(K_2 \mid K_1) && (4.10) \\ &= P(K_2) \cdot P(K_1 \mid K_2) && (4.11) \end{aligned}$$

De twee zojuist geformuleerde regels noemt men de *algemene productregel(s)*.

Voorbeeld 15

Uit een volledig spel kaarten worden 2 kaarten getrokken. De trekking geschiedt zonder teruglegging, dat wil zeggen dat de eerstgetrokken kaart niet in het spel wordt teruggelegd alvorens de tweede kaart wordt getrokken. Hoe groot is de kans dat beide kaarten harten zijn?

Oplossing

We dienen in dit geval te bedenken dat, wanneer de eerstgetrokken kaart een harten is, het spel daarna nog $52 - 1 = 51$ kaarten bevat waarvan er $13 - 1 = 12$ harten zijn. Met $K_1 = \{\text{de eerste kaart is harten}\}$ en $K_2 = \{\text{de tweede kaart is harten}\}$ leidt toepassing van rekenregel (4.10) tot:

P(de eerste kaart is harten én de tweede kaart is harten) =

$P(K_1 \cap K_2) = P(K_1) \cdot P(K_2 \mid K_1) =$

P(de eerste kaart is harten) $\cdot$ P (de tweede kaart is harten | de eerste kaart is harten) =

$\frac{13}{52} \cdot \frac{12}{51} = \frac{1}{17}$.

4.5.5 Afhankelijkheid en onafhankelijkheid

Wanneer het optreden van gebeurtenis K_1 geen invloed heeft op de kans van optreden van gebeurtenis K_2 – dat wil zeggen wanneer $P(K_2)$ hetzelfde blijft ongeacht of K_1 nu wel of niet is opgetreden – kunnen we voor $P(K_2|K_1)$ gewoon $P(K_2)$ schrijven. In het geval dat

$P(K_1|K_2) = P(K_1) \neq 0$ én $P(K_2|K_1) = P(K_2) \neq 0$ zeggen we dat de gebeurtenissen K_1 en K_2 *onafhankelijk* zijn.
Wanneer het optreden van gebeurtenis K_2 wel beïnvloed wordt door het optreden van gebeurtenis K_1 (en/of omgekeerd), worden K_1 en K_2 *afhankelijke* gebeurtenissen genoemd. In dat geval is $P(K_1|K_2) \neq P(K_1)$ en/of $P(K_2|K_1) \neq P(K_2)$. Bij afhankelijke gebeurtenissen gebruiken we rekenregel (4.10) of (4.11) (de algemene productregel(s)). Bij onafhankelijke gebeurtenissen gebruiken we de zogenaamde *speciale productregel.*

4.5.6 De speciale productregel

Stel: de gebeurtenissen K_1 en K_2 zijn onafhankelijke gebeurtenissen. Rekenregel (4.10) of (4.11) gaat dan over in rekenregel (4.12), de zogenaamde *speciale productregel.*

Stelling 6

Wanneer K_1 en K_2 onafhankelijke gebeurtenissen zijn, dat wil zeggen wanneer $P(K_1|K_2) = P(K_1) \neq 0$ en $P(K_2|K_1) = P(K_2) \neq 0$, geldt:

$$P(K_1 \cap K_2) = P(K_1) \cdot P(K_2) \tag{4.12}$$

Voorbeeld 16

De kans om met twee dobbelstenen 2 'zessen' te gooien, is op twee manieren te berekenen:

a. Direct, met de klassieke definitie van het kansbegrip (zie hiervoor voorbeeld 1).

b. Met de speciale productregel:
Met K_1 = {6 ogen in de eerste worp} en K_2 = {6 ogen in de tweede worp} kunnen we stellen dat K_1 en K_2 onafhankelijk zijn, want de kans om bij de tweede worp een bepaald aantal ogen te gooien, wordt niet beïnvloed door het resultaat van de eerste worp. We mogen dus rekenregel (4.12) toepassen en vinden dan:
P(6 ogen in de eerste worp en 6 ogen in de tweede worp) =
$P(K_1 \cap K_2) =$
$P(K_1) \cdot P(K_2) =$
P (6 ogen in de eerste worp) $\cdot$ P (6 ogen in de tweede worp) = $\frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$.

4.6 Combinatoriek

In de volgende paragrafen zullen we de lezer voorbereiden op het maken van iets ingewikkelder kansvraagstukken en het bestuderen van kansverdelingen (hoofdstuk 5). Daartoe is het handig om enige kennis te hebben van de begrippen *permutaties, variaties en combinaties.*

4.6.1 Permutaties

Van de cijfers 1 en 2 kunnen we op twee manieren een getal van twee cijfers maken: 12 en 21. van de cijfers 1, 2 en 3 kan op 6 manieren een getal van 3 cijfers gemaakt worden: 123, 132, 213, 231, 312 en 321. Stel nu dat we de beschikking hebben over 5 verschillende cijfers: 1, 2, 3, 4 en 5. Op hoeveel manieren kunnen we van deze 5 cijfers een getal van 5 cijfers maken? Het aantal manieren is te groot om uit te schrijven maar kan gemakkelijk berekend worden: het cijfer waarmee het getal begint kan op 5 verschillende manieren gekozen worden. Voor het tweede cijfer kan dan gekozen worden uit de 4 resterende cijfers. De eerste twee cijfers kunnen dus op $5 \times 4 = 20$ verschillende manieren gekozen worden. Voor het derde cijfer resteren nog 3 keuzemogelijkheden, zodat het aantal manieren waarop de eerste drie cijfers gekozen kunnen worden, gelijk is aan $5 \times 4 \times 3 = 60$. Voor het vierde cijfer resteren nog 2 keuzemogelijkheden en voor het vijfde cijfer nog slechts 1. Met de 5 verschillende cijfers kan dus op $5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$ verschillende manieren een getal van 5 cijfers gemaakt worden. We zeggen nu dat de 5 verschillende cijfers op 120 manieren gerangschikt of gepermuteerd kunnen worden. Het aantal *permutaties* van de 5 verschillende cijfers bedraagt 120.

Opmerking

Het product $5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$ kan geschreven worden als 5! (lees: 5-faculteit).
Het begrip faculteit is bekend uit de wiskunde. Hierin wordt $n!$ gedefinieerd als:
$n! = (n)(n-1)(n-2)...(3)(2)(1)$ voor elk natuurlijk getal n, waarbij 1! gedefinieerd is als 1 en 0! eveneens als 1.
Een belangrijke eigenschap van faculteiten is dat $\frac{k!}{(k-1)!} = k$, dus bijvoorbeeld $\frac{7!}{6!} = 7$.

We kunnen nu de volgende definitie formuleren:

Definitie

Men kan een groep van n verschillende elementen op $P_n = n!$ manieren rangschikken of permuteren. Elke rangschikking heet een permutatie.

Stel nu dat een getal van 5 cijfers uitsluitend bestaat uit 3 nullen en 2 enen (zo'n getal heet een binair getal), bijvoorbeeld 10100. Zo'n getal kan als volgt worden gemaakt (let op de systematiek):

11000	10100	10010	10001	01100
01010	01001	00110	00101	00011

We tellen tien manieren om 3 nullen en twee enen te rangschikken tot een getal van 5 cijfers. Wanneer we de drie nullen zouden vervangen door verschillende symbolen, bijvoorbeeld A, B en C (zodat 11000 wordt 11ABC), wordt het aantal mogelijke rangschikkingen 3! keer

groter. Immers, de drie letters A, B en C kunnen in elk van de tien binaire getallen op 3! manieren worden gepermuteerd. We kunnen dus op $10 \times 3! = 60$ manieren een 'getal' vormen met de symbolen A, B, C en de twee enen.
Vervangen we nu ook de twee enen door verschillende symbolen, bijvoorbeeld door de letters X en Y (voorbeeld: XAYBC, maar ook YAXBC), dan neemt het aantal rangschikkingen met een factor 2! toe tot $60 \times 2! = 120$. Dit klopt, want er is een 'getal' ontstaan dat bestaat uit 5 verschillende symbolen en we wisten al dat zo'n getal op 5! manieren is te maken.
Terugredeneren leidt nu tot de conclusie, dat een getal van 5 cijfers dat opgebouwd is uit 3 nullen en 2 enen op $\frac{5!}{3!2!} = 10$ manieren is te maken.
We kunnen nu de volgende definitie geven:

Definitie

Wanneer een groep van N elementen is verdeeld in k groepjes van n_1 gelijke elementen, n_2 gelijke elementen, n_3 gelijke elementen, enzovoorts (zodanig dat $n_1 + n_2 + n_3 + \ldots + n_k = N$) dan is het aantal rangschikkingen (permutaties) van deze N elementen gelijk aan

$$P^N_{n_1,n_2,n_3,\ldots,n_k} = \frac{N!}{n_1!n_2!n_3!\ldots n_k!} \tag{4.13}$$

Voorbeeld 17

Op hoeveel verschillende manieren kan een gezin:

a. met 4 kinderen 2 jongens en 2 meisjes tellen?
b. met 8 kinderen 4 jongens en 4 meisjes tellen?

Oplossing

a. Met $N = 4$ (kinderen), $n_1 = 2$ (jongens) en $n_2 = 2$ (meisjes) vinden we volgens de definitie van permutaties dat het gevraagde aantal gelijk is aan $P^4_{2,2} = \frac{4!}{2!2!} = 6$.
b. Met $N = 8$, $n_1 = 4$ en $n_2 = 4$ vinden we: $P^8_{4,4} = \frac{8!}{4!4!} = 70$.

4.6.2 Variaties

Stel dat we de beschikking hebben over 5 verschillende cijfers, bijvoorbeeld 1, 2, 3, 4 en 5. Uit deze groep van 5 cijfers nemen we er drie en vormen daarmee een getal van 3 cijfers. Op hoeveel manieren kunnen we dit doen? Het verschil met de situatie die hierboven is beschouwd, is dat we niet alle cijfers gebruiken maar slechts een deel ervan. Wel zijn de drie te kiezen cijfers verschillend. We moeten daarom bedenken dat het getal dat gevormd wordt door bijvoorbeeld de cijfers 2, 3 en 5 op 3! manieren kan geschieden.
Voor het eerste cijfer van het getal kunnen we kiezen uit 5 mogelijkheden. Voor het tweede cijfer resteren nog 4 mogelijkheden en voor het derde cijfer hebben de keus uit de resterende 3 cijfers. In totaal kunnen we dus op $5 \cdot 4 \cdot 3 = 60$ verschillende manieren een getal van drie verschillende cijfers (te kiezen uit 5 verschillende cijfers).

Bedenken we dat $5 \cdot 4 \cdot 3$ geschreven kan worden als $\dfrac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1} = \dfrac{5!}{2!} = \dfrac{5!}{(5-3)!}$, dan hebben we voor het aantal manieren waarop we uit 5 beschikbare cijfers er 3 kunnen kiezen, een schrijfwijze gevonden, waarin zowel het aantal beschikbare elementen (5) als het aantal daaruit te kiezen elementen (3) wordt genoemd. Aldus komen we tot de volgende algemene definitie:

Definitie

Het aantal manieren waarop k verschillende elementen, uit een groep van n verschillende elementen $(0 \leq k \leq n)$ kan worden gerangschikt, is :

$$V_k^n = \frac{n!}{(n-k)!} \tag{4.14}$$

V_k^n wordt uitgesproken als 'V k uit n' = het aantal *variaties* van k elementen uit een groep van n elementen.

In het volgende voorbeeld geven we een toepassing van het begrip variaties in de kansrekening.

Voorbeeld 18

De 10 paarden die meedoen aan een harddraverij zijn genummerd van 1 tot en met 10. Bij een 'trio' gaat het erom de nummers van de drie eerst aankomende paarden in volgorde van aankomst te voorspellen. Hoe groot is, onder de aanname dat voor ieder paard de winstkans even groot is, de kans op een goede voorspelling, wanneer men de drie nummers willekeurig kiest?

Oplossing

Bedenk dat niet alleen de nummers van de drie eerst aankomende paarden voorspeld moeten worden maar juist ook de volgorde van die nummers. Daarom moeten we het aantal manieren waarop we uit de 10 nummers 3 nummers kunnen kiezen, berekenen volgens de definitie van het begrip variaties. We vinden dan $V_3^{10} = \dfrac{10!}{(10-3)!} = 720$. Er is maar één volgorde winnend. De gevraagde kans is dan volgens de klassieke kansdefinitie gelijk aan $\frac{1}{720}$.

Wanneer de volgorde niet van belang is, krijgen we te maken met *combinaties* in plaats van variaties.

4.6.3 Combinaties

Stel dat we uit een projectgroep van vijf personen (bijvoorbeeld A, B, C, D en E) er drie willen kiezen in het 'bestuur' van de projectgroep. Op hoeveel manieren kan dat?

Dit voorbeeld lijkt veel op een vorig voorbeeld (na de definitie van permutaties) waarbij we uit een groep van 5 verschillende cijfers er 3 gekozen hebben om daarmee een getal

van drie cijfers te vormen. Het verschil is dat bij de keuze van elk drietal cijfers ook de onderlinge permutaties van die cijfers moesten worden meegeteld (235 is een ander getal dan 352), terwijl bij de keuze van drie personen uit de groep van 5 personen de volgorde niet van belang is. Deze volgorde zou wel van belang zijn als we ons afvroegen op hoeveel manieren uit de 5 personen een voorzitter, een secretaris en een penningmeester (dus drie verschillende functionarissen) gekozen zouden kunnen worden. In dat geval is bijvoorbeeld de keuze A (voorzitter), B (secretaris), C (penningmeester) een andere dan B (voorzitter), C (secretaris) en A (penningmeester). In het laatste geval (wanneer de volgorde dus wel van belang is) hebben we het aantal variaties van 3 uit 5, dus $\frac{5!}{(5-3)!} = 60$. Maar wanneer we alleen geïnteresseerd zijn in de drie personen die het bestuur vormen en niet in hun functies (= 'volgorde') krijgen we een veel kleiner aantal rangschikkingen. Wanneer we het aantal rangschikkingen van 3 uit 5 *ongeacht de volgorde* op x stellen, moet $x \cdot 3!$ het aantal variaties (*geacht de volgorde*) opleveren. Dus in dit geval is

$$x = \frac{\text{aantal variaties}}{3!} = \frac{\frac{5!}{(5-3)!}}{3!} = \frac{5 \cdot 4 \cdot 3}{6} = 10$$

Het betreft de volgende 10 rangschikkingen:

ABC ABD ABE ACD ACE ADE BCD BCE BDE CDE

Ter onderscheid met het begrip variatie (mét volgorde) gebruiken we het begrip *combinatie*, wanneer de volgorde binnen een rangschikking er niet toe doet.
Aldus kunnen we definiëren:

Definitie
Voor een groep van k elementen uit een groep van n elementen ($0 \leq k \leq n$) is – wanneer de volgorde er niet toe doet, maar alleen de keuze – het aantal van elkaar te onderscheiden rangschikkingen (die dan de combinaties worden genoemd) gelijk aan:

$$C_k^n = \frac{n!}{k!(n-k)!} \tag{4.15}$$

C_k^n wordt uitgesproken als 'C k uit n' = het aantal combinaties van k elementen uit een groep van n elementen. C_k^n wordt ook wel *binomiaalcoëfficiënt* genoemd, met als notatie $\binom{n}{k}$.
Een andere invalshoek om naar binomiaalcoëfficiënten te kijken zullen we illustreren met het volgende voorbeeld.
De vraag 'op hoeveel manieren kan uit een groep van 5 personen een bestuur van 3 personen gekozen worden' kan vertaald worden naar 'op hoeveel manieren kan uit een groep van 5

personen een groepje van 3 personen worden gekozen (dus 2 niet gekozen)'. Om dit uit te rekenen, gebruiken we de formule voor $P^5_{2,3} = \dfrac{5!}{2!3!} = 10$. Het aantal combinaties is dus niets anders dan het aantal permutaties van twee verschillende groepen elementen namelijk de 3 gekozen personen en de 2 niet-gekozen personen.
In het volgende voorbeeld laten we een toepassing van het begrip combinaties zien in de kansrekening.

Voorbeeld 19

Hoe groot is de kans dat een gezin met 4 kinderen 2 jongens (en dus 2 meisjes) telt, wanneer we aannemen dat de geboortekans zowel voor een jongen als voor een meisje gelijk is aan $\frac{1}{2}$?

Oplossing

Gezien de gestelde vraag is het geboortenummer of de leeftijdsvolgorde van de 4 kinderen niet van belang. Alleen het geslacht is van belang. Daarom geldt voor het aantal manieren waarop een gezin met 4 kinderen 2 jongens (en dus 2 meisjes) kan tellen, dat dit gelijk is aan $C^4_2 = \dfrac{4!}{2!2!} = 6$.
Alle 6 de volgordes zijn even waarschijnlijk met kans $\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{16}$ (speciale productregel) en sluiten elkaars optreden uit. De somregel voor elkaar uitsluitende gebeurtenissen mag daarom toegepast worden. De gevraagde kans is dus: $P(2 \text{ jongens en } 2 \text{ meisjes}) = 6 \cdot \left(\frac{1}{2}\right)^4 = \frac{3}{8}$.

4.7 Het oplossen van kansvraagstukken

Het oplossen van kansvraagstukken en van statistische problemen, waarbij de theorie van de kansrekening moet worden toegepast, is in het algemeen niet eenvoudig. Dit komt omdat het vaak niet direct duidelijk is welke rekenregel moet worden toegepast. We zullen daarom enkele aanwijzingen geven waarmee het oplossen van kansvraagstukken enigszins gesystematiseerd kan worden en daardoor eenvoudiger wordt:

1. Is er sprake van een 'én'-situatie, dat wil zeggen wordt er gevraagd naar de kans dat zich de gebeurtenis K_1 én de gebeurtenis K_2 zullen voordoen, dan moet een *productregel* worden toegepast. Vaak is het zo dat het woordje 'en' verdekt aanwezig is. Wanneer bijvoorbeeld gevraagd wordt naar de kans dat in een steekproef van 5 willekeurige personen er twee vrouwen zijn, dan wordt in feite bedoeld: twee vrouwen én drie mannen.

2. Is eenmaal vastgesteld dat een der productregels moet worden toegepast, ga dan na of de gebeurtenissen wel of niet onafhankelijk zijn. Zijn ze onafhankelijk, gebruik dan de *speciale productregel*; zijn ze afhankelijk, gebruik dan de *algemene productregel*.

Bij steekproeven met teruglegging is sprake van onafhankelijkheid, bij steekproeven zonder teruglegging is daarentegen sprake van afhankelijkheid. Ga dit na!

3. Is er sprake van een 'óf'-situatie, dat wil zeggen wordt er gevraagd naar de kans dat zich de gebeurtenis K_1 óf de gebeurtenis K_2 zal voordoen, dan moet een *optelregel* worden toegepast. Soms is niet direct duidelijk dat het om een óf-situatie gaat. Wanneer bijvoorbeeld gevraagd wordt de kans om in een steekproef van 10 stuks hoogstens 2 producten met het predikaat 'van slechte kwaliteit' aan te treffen, is er sprake van een òf-situatie. De vraag houdt namelijk in: 0 of 1 of 2 producten van slechte kwaliteit. Merk op dat hier ook nog eens het woordje 'en' in verstopt zit. Immers, bijvoorbeeld 2 slechte producten betekent in feite 2 slechte én 8 goede producten.
4. Heeft men eenmaal vastgesteld dat een optelregel moet worden toegepast, ga dan na of de gebeurtenissen elkaar wel of niet uitsluiten. Sluiten ze elkaars optreden uit (óf de ene, óf de andere, maar niet allebei tegelijk) gebruik dan de speciale optelregel. Sluiten ze elkaars optreden niet uit (ze kunnen tegelijk optreden) gebruik dan de algemene optelregel. In de praktijk zullen we vaker te maken hebben met elkaar uitsluitende gebeurtenissen. Een eenvoudig voorbeeld: wanneer gevraagd wordt naar de kans om hoogstens twee slechte producten in een steekproef van 5 producten aan te treffen, wordt bedoeld de elkaar uitsluitende gebeurtenissen 0, 1 of 2 slechte producten.
5. Het verdient aanbeveling bij elk kansvraagstuk na te gaan of het besparing van rekenwerk kan opleveren door de *complementregel* toe te passen. Het is vaak een kwestie van intuïtie of men daartoe besluit. Soms ligt het duidelijk voor de hand. Wanneer bijvoorbeeld gevraagd wordt naar de kans om in een steekproef van 50 producten meer dan 1 slecht product aan te treffen, kan men deze kans het beste met de complementregel uitrekenen: 'meer dan 1' heeft als complement 'hoogstens 1', dus 0 of 1.
6. Indien er sprake is van $m \times n$ gebeurtenissen van m verschillende kenmerken met n verschillende kenmerken, is het handig om een $m \times n$-tabel te maken (zie voorbeeld 13).
7. Met combinatoriek (permutaties, variaties en combinaties) kan in veel gevallen het telwerk vergemakkelijkt worden. Wanneer het aantal (elkaar uitsluitende) mogelijkheden moet worden berekend, kunnen vaak de formules (4.14) of (4.15) worden toegepast. Op de begrippen uit de combinatoriek komen we in het volgende hoofdstuk uitvoerig terug (zie bij binomiale kansverdeling en hypergeometrische kansverdeling).
8. Soms is het handig om een *kansboom* te maken. We geven een voorbeeld. Er wordt driemaal geworpen met een onzuivere munt. De kans dat 'kop' verschijnt is $\frac{2}{3}$ en de kans dat 'munt' verschijnt is $\frac{1}{3}$. Alle mogelijke gebeurtenissen kunnen met de bijbehorende kansen in een kansboom worden weergegeven.

Uit de kansboom kunnen we bijvoorbeeld opmaken dat de kans dat er in drie worpen twee maal 'kop' (dus éénmaal 'munt') verschijnt, gelijk is aan de som van de kansen $P(K_1 \cap K_2 \cap M_3)$, $P(K_1 \cap M_2 \cap K_3)$ en $P(M_1 \cap K_2 \cap K_3)$, dus $\frac{12}{27}$.

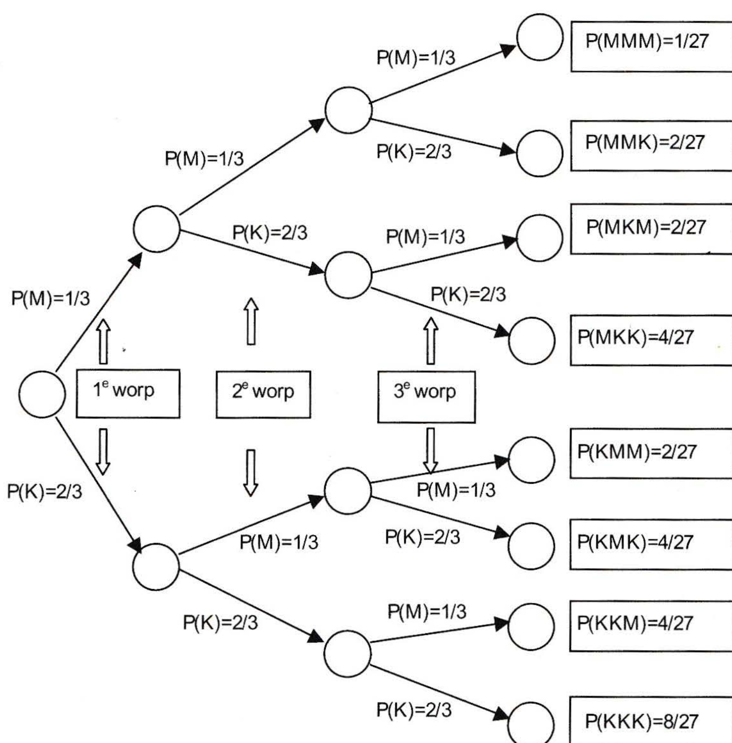


Fig. 4.6 Een kansboom

Opgaven

1. Een groep van tien mensen is als volgt samengesteld:

persoon	1	2	3	4	5	6	7	8	9	10
geslacht	*M*	*V*	*M*	*V*	*V*	*M*	*V*	*V*	*M*	*V*
leeftijd	31	44	56	34	22	67	42	23	51	46
aantal kinderen	0	2	3	1	0	2	0	0	2	3

Uit deze groep wordt aselect (dus volkomen willekeurig) iemand op basis van de kenmerken geslacht (M/V), leeftijd (50 of ouder/ onder de 50) en op het hebben van kinderen (wel/geen) geselecteerd.

a. Bepaal:
- $P(M)$
- $P(50 \text{ of ouder})$
- $P(\text{wel kinderen})$
- $P(M \mid \text{wel kinderen})$

- $P(\text{wel kinderen}| M)$
- $P(V |\text{onder de 50})$
- $P(\text{onder de 50}| V)$

Pas, indien mogelijk, de kansregels toe op de volgende vragen.

b. Hoe groot is de kans dat de geselecteerde persoon een man is zonder kinderen?

c. Wanneer gegeven is dat de geselecteerde persoon een man is, hoe groot is dan de kans dat hij boven de 50 is én kinderen heeft?

d. Hoe groot is de kans dat de geselecteerde persoon een vrouw is, óf iemand van boven de 50?

Er worden nu twee personen geselecteerd.

e. Hoe groot is de kans dat beide personen man zijn?

f. Hoe groot is de kans dat een van de personen man is?

g. Hoe groot is de kans dat beide personen zowel boven de 50 zijn als kinderen hebben?

2. Er wordt met drie zuivere dobbelstenen gegooid.

a. Hoe groot is de kans om in totaal minstens 5 ogen te gooien?

b. Hoe groot is de kans om tweemaal een zes (dus eenmaal geen zes) te gooien?

3. Men gooit met twee zuivere dobbelstenen en bepaalt de som A en het product B van de beide aantallen ogen.

a. Hoe groot is de kans dat A groter is dan 7?

b. Hoe groot is de kans dat B oneven is?

c. Hoe groot is de kans dat A groter dan 7 en B oneven is?

d. Hoe groot is de kans dat A groter dan 7 of B oneven is?

4. Men trekt zonder teruglegging twee kaarten uit een volledig kaartspel.

a. Hoe groot is de kans op twee harten?

b. Hoe groot is de kans op een harten en een schoppen?

c. Hoe groot is de kans op twee harten of twee ruiten?

5. Op grond van het verleden verwacht iemand dat er een kans van 70% bestaat dat de AEX-index volgend jaar zal stijgen en een kans van 25% dat de AEX-index zal dalen (er is dus een kans van 5% dat de AEX-index gelijk blijft). Verwacht wordt (ook op grond van het verleden) dat de aandelenkoers van de firma Y met een kans van 80% zal stijgen, als de AEX-index stijgt. Als de AEX-index gelijk blijft, wordt verwacht dat de aandelenkoers van firma Y met een kans van 50% zal stijgen, met een kans van 30% gelijk blijft (en dus met een kans van 20% zal dalen). En als de AEX-index volgend jaar daalt, wordt verwacht dat de aandelen van firma Y met een kans van 80% in waarde zullen dalen, met een kans van 10% in waarde zullen stijgen (en dus met een kans van 10% gelijk blijven).

Wat is de kans dat de aandelenkoers van firma Y volgend jaar zal stijgen?

6. Een machine produceert een bepaald type kogellagers. Uit ervaring is bekend dat in 8 van de 10 gevallen de machine goed is ingesteld. Wanneer de machine goed is afgesteld, wordt 90% correct gefabriceerd. Wanneer de machine niet juist is afgesteld, wordt 30% van de kogellagers niet correct geproduceerd.

a. Hoe groot is de kans dat een kogellager correct wordt gefabriceerd?
Na een zekere set-up wordt de eerste geproduceerde kogellager correct gefabriceerd.

b. Hoe groot is de voorwaardelijke kans dat de machine goed was afgesteld?

7. Een interviewbureau wordt ingeschakeld om een bepaald product telefonisch te promoten. De doelgroep bestaat uit personen van 40 jaar en ouder, die gemiddeld meer dan 3 uur per dag naar de televisie kijken. Bekend is dat de populatie telefoonbeantwoorders voor 60% uit personen van 40 jaar en ouder bestaat. Ook is bekend dat 20% van de telefoonbeantwoorders gemiddeld meer dan 3 uur per dag naar de televisie kijkt. Hoeveel mensen moeten worden opgebeld om te bereiken dat naar verwachting van de doelgroep minstens 1000 personen worden bereikt? Neem aan dat de leeftijd en het aantal uren per dag televisie kijken geen invloed op elkaar hebben.

8. In een ijzerwarenfabriek staan drie machines (A, B en C), elk goed voor respectievelijk 20%, 30% en 50% van de totale schroevenproductie. Van de productie van machine A voldoet 3% niet aan de gestelde kwaliteitsnormen; voor machine B is dit 5% en voor machine C 10%. Uit de totale productie wordt willekeurig een schroef genomen.

a. Hoe groot is de kans dat de schroef niet aan de gestelde kwaliteitseis voldoet?

b. De schroef blijkt niet aan de gestelde kwaliteitsnorm te voldoen. Hoe groot is de kans dat de schroef door machine C vervaardigd is?

9. In de wielersport gebruikt 10% van de topsporters doping. Wanneer iemand doping gebruikt heeft, wordt dit in 95% van de gevallen bij een dopingcontrole ontdekt (positief). Maar wanneer iemand geen doping heeft gebruikt, geeft de dopingcontrole slechts in 92% van de gevallen een correcte uitslag.

a. Wat is de kans dat een willekeurige wielrenner doping gebruikt en ook positief bevonden wordt?

b. Hoe groot is de kans dat bij de dopingcontrole een willekeurige wielrenner positief bevonden wordt?

c. Een wielrenner wordt positief bevonden. Hoe groot is de kans dat hij toch geen doping heeft gebruikt?

10. De schutters Aad, Ben en Chris hebben een verschillend scoringspercentage: Aad schiet 7 van de 10 keer in de roos, Ben 5 keer van de 10 keer en Chris 9 van de 10 keer. Aad, Ben en Chris zijn aan het trainen. Maar Aad schiet 2 keer zoveel als Ben en 3 keer zoveel als Chris.

a. Hoe groot is de kans dat een willekeurig schot de roos treft?

b. Een schot treft de roos. Hoe groot is de kans dat het schot van Aad afkomstig is?

11. Van de 1500 tv-apparaten die in een jaar tijd in het servicecentrum van de firma Sonata ter reparatie werden aangeboden voordat de garantietermijn van 1 jaar verstreken was, is bekend dat er 600 geassembleerd waren in de fabriek te München en 900 in de fabriek te Seoel. Na uitgebreid onderzoek bleken de apparaten ofwel door materiaalfouten, ofwel door constructiefouten defect te zijn geraakt en wel volgens onderstaand overzicht:

	Seoel	München	Totaal
materiaalfout	600	500	1100
constructiefout	300	100	400
Totaal	900	600	1500

a. Wat is de kans dat een willekeurig tv-toestel (uit de 1500 tv-toestellen) een materiaalfout vertoonde?

b. Wat is de kans dat een willekeurig tv-toestel in Seoel gefabriceerd werd?

c. Wat is de kans dat een willekeurig tv-toestel een materiaalfout had én in Seoel gefabriceerd werd?

d. Zijn de gebeurtenissen 'materiaalfout' en 'in Seoel gefabriceerd' onafhankelijk of niet?

e. Wat is de kans dat een toestel met een materiaalfout in Seoel werd gefabriceerd? Bereken deze kans op twee manieren: direct uit de tabel en met behulp van de kansregels.

12. Een projectgroep bestaat uit 8 personen: 6 mannen en 2 vrouwen.

a. Op hoeveel manieren kunnen een voorzitter en een secretaris gekozen worden?

b. Op hoeveel manieren kunnen een voorzitter en een secretaris gekozen worden onder de voorwaarde dat ze niet van hetzelfde geslacht zijn?

c. Op hoeveel manieren kunnen een voorzitter en een secretaris gekozen worden onder de voorwaarde dat de voorzitter vrouw is en de secretaris man?

13. a. Op hoeveel manieren kan men uit de cijfers 1, 2, 3, 3, 4, 4, 5 een getal van zeven cijfers maken?

b. Op hoeveel manieren kan met de letters van het woord ABACADABRA een lettercombinatie gemaakt worden van 10 letters?

14. De play-offs van de nationale basketbalcompetitie worden gespeeld in twee poules van elk 4 teams. De nummers 1 en 2 van beide poules spelen met elkaar in een finalepoule van 4 teams. De nummers 1, 2 en 3 uit de finalepoule krijgen een medaille (goud, zilver, brons). Hoeveel mogelijke rangschikkingen zijn er voor de medaillewinnaars?

15. Bij de lotto worden uit 45 balletjes, genummerd van 1 t/m 45, er 6 getrokken. Hoe groot is de kans dat iemand alle nummers goed raadt? En hoe groot is de kans dat iemand 5 nummers goed raadt?

16. Hoe groot is de kans dat iemand bij een multiple-choice test van 10 vragen met elk 3 mogelijke antwoorden op goed geluk precies 3 goede antwoorden (dus 7 foute antwoorden) aankruist?

17. In een kroeg zitten 20 stamgasten. Hoe groot is de kans dat 2 of meer stamgasten op dezelfde dag jarig zijn? Tip: Bereken deze kans met behulp van de complementregel.

18. In een grote partij producten zitten 5% defecte exemplaren. Iemand neemt een steekproef van 10 producten.

a. Hoe groot is de kans dat er geen defecte exemplaren bij zitten?
b. Hoe groot is de kans dat er hoogstens 1 defect exemplaar bij zit?
c. Hoe groot is de kans dat er meer dan 2 defecte exemplaren bij zitten?

19. In een productieserie van 20 producten zitten 6 defecte exemplaren. Iemand neemt uit deze serie een steekproef van 5 producten.

a. Hoe groot is de kans dat er geen defecte exemplaren bij zitten?
b. Hoe groot is de kans dat er hoogstens 1 defect exemplaar bij zit?
c. Hoe groot is de kans dat er meer dan 2 defecte exemplaren bij zitten?

20. Bij een onderzoek naar het aantal pogingen dat men nodig heeft om het praktisch rijexamen te halen, is naar voren gekomen dat 30% het examen in één keer haalt, 20% het de tweede keer haalt, 10% succes bij de derde poging heeft en de rest meer dan 3 pogingen nodig heeft. Een groep van 4 vrienden wil het praktisch rijexamen halen.

a. Hoe groot is de kans dat alle vier in één keer slagen?
b. Hoe groot is de kans dat één vriend het examen in één keer haalt, één het examen in twee keer haalt, één het in drie keer en één in meer dan drie keer?
c. Hoe groot is de kans dat twee vrienden het examen in één keer halen en de andere twee in meer dan één keer?

5 Discrete kansverdelingen

5.1 Inleiding

In hoofdstuk 3 hebben we een aantal begrippen uit de beschrijvende statistiek besproken. Zo kan men een reeks van waarnemingsuitkomsten karakteriseren door een maatstaf voor de ligging (bijvoorbeeld rekenkundig gemiddelde) en een maatstaf voor de spreiding (onder andere variantie). Verder hebben we de waarnemingsuitkomsten overzichtelijk weergegeven door middel van een frequentietabel. In hoofdstuk 4 hebben we kennisgemaakt met allerlei situaties, waarbij het toeval (=kans) een belangrijke rol speelt. In dit hoofdstuk gaan we nu de begrippen uit de beschrijvende statistiek ook toepassen op die situaties waarbij toeval een rol speelt. We spreken van een *kansverdeling* indien bij een 'kansexperiment' de mogelijke gebeurtenissen met betrekking tot een bepaald kenmerk met de bijbehorende kansen kunnen worden genoteerd. Zoals in hoofdstuk 2 reeds is opgemerkt, noemt men het kenmerk waarop de kansen betrekking hebben een *kansvariabele*. De waarde of uitkomst van het kenmerk hangt immers van het toeval af. De kansverdeling behorend bij een bepaalde kansvariabele is vergelijkbaar met een frequentietabel, waar de gemeten uitkomsten en de bijbehorende frequentie zijn opgeschreven. Net als een frequentieverdeling kan een kansverdeling gekarakteriseerd worden door parameters voor de ligging en de spreiding. Als parameter voor de ligging wordt bijna altijd de *verwachting* of de *verwachtingswaarde* (= gemiddelde) gebruikt. Voor de spreiding gebruiken we bij kansverdelingen net als bij frequentieverdelingen de begrippen *variantie* en *standaardafwijking*.
We onderscheiden twee typen kansverdelingen, namelijk de *discrete* en de *continue* kansverdeling. Een *discrete kansverdeling* ontstaat als de kansvariabele *discreet* (oftewel *discontinu)* is. Dit is bijvoorbeeld het geval als de uitkomst van een gebeurtenis (of kansexperiment) is te tellen (bijvoorbeeld het aantal foutieve exemplaren in een bepaalde partij, het aantal ongelukken op een bepaalde plaats, het aantal ziektegevallen, enzovoorts). In dit hoofdstuk bespreken we de voor de dagelijkse praktijk belangrijkste discrete kansverdelingen, zoals de *binomiale verdeling* en de *Poisson-verdeling*.

Een *continue kansverdeling* ontstaat als de te meten variabele continu is (bijvoorbeeld lengte, gewicht, concentratie of temperatuur). De verreweg belangrijkste continue kansverdeling is de *normale* verdeling; hierop gaan we in hoofdstuk 6 verder in.

5.2 Discrete kansverdeling

Zoals we in de inleiding al hebben gezien, zijn er twee typen van kansverdelingen, één waarbij de kansvariabele discreet (discontinu) is en één waarbij de kansvariabele continu is. In deze paragraaf beginnen we met een voorbeeld van een discrete kansverdeling.

Voorbeeld 1

We gooien met twee dobbelsteen en definiëren de kansvariabele K = som van het aantal ogen dat geworpen wordt. Alle mogelijke uitkomsten k van dit experiment kunnen we opschrijven met de erbij behorende kans. We krijgen dan de kansverdeling van de kansvariabele K.

Tabel 5.1 Mogelijke waarden van K=som van het aantal ogen

		dobbelsteen 1					
		1	**2**	**3**	**4**	**5**	**6**
dobbelsteen 2	**1**	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

Uit de tabel kan men afleiden dat de kansvariabele K in totaal 36 uitkomsten heeft, die echter niet allemaal verschillend zijn. De uitkomst 2 komt slechts één keer voor, de kans op de uitkomst 2 in dit experiment is dus $\frac{1}{36}$. Dit schrijven we als volgt: $P(K = 2) = \frac{1}{36}$. De kans op de uitkomst van bijvoorbeeld 7, die 6 keer voorkomt, is dus $P(K = 7) = \frac{6}{36}$. In dit voorbeeld zien we dat de kansvariabele K alleen gehele waarden kan aannemen, de kansverdeling van K is dus een *discrete* kansverdeling .

De kansverdeling van K is:

k	2	3	4	5	6	7	8	9	10	11	12	$\sum$
$P(K = k)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	$\frac{36}{36} = 1$

Opmerking

Meestal gebruikt men voor een discrete kansvariabele het symbool K en voor een continue kansvariabele het symbool X. Zoals in hoofdstuk 2 reeds is opgemerkt, wordt met K en X de naam van de variabele bedoeld. De naam van een variabele schrijven we altijd met een hoofdletter. Met k en x bedoelen we de *waarde* van de variabele, geschreven met een kleine letter. Verder voeren we het symbool $f(x)$ respectievelijk $f(k)$ in, dat in wiskunde bekend staat als 'functie van x, respectievelijk k'. In de kansrekening spreken we echter van *kansfunctie* $f(k)$ als K een discrete kansvariabele is en van een *kansdichtheid* $f(x)$ als X een continue kansvariabele is.

Definitie

Voor de kansfunctie van een discrete kansvariabele geldt dat: $f(k) = P(K = k)$, voor iedere waarde k, die K kan aannemen.

In voorbeeld 1 is dus: $f(k) = P(K = 12) = \frac{1}{36}$.
De kansfunctie geeft de *individuele* kansen per gebeurtenis aan.
Verder kennen we in de kansrekening nog het begrip *verdelingsfunctie*.
De verdelingsfunctie van K wordt weergegeven door het symbool $F(k)$.
Voor een discrete kansvariabele K verstaan we onder $F(k)$:
'De kans op de gebeurtenis dat K kleiner of gelijk is aan de waarde k.'
In voorbeeld 1 is:
$F(5) = P(K \leq 5) =$
$P(K = 2) + P(K = 3) + P(K = 4) + P(K = 5) = \frac{1}{36} + \frac{2}{36} + \frac{3}{36} + \frac{4}{36} = \frac{10}{36}$
De verdelingsfunctie $F(k)$ is dus op te vatten als de *cumulatieve* kans*cumulatieve kans* van de kansvariabele K.

Definitie

Voor een discrete kansvariabele K geldt dat de verdelingsfunctie $F(k)$ gelijk is aan: $F(k) = P(K \leq k) = \sum_{i \leq k} P(K = i) = \sum_{i \leq k} f(i)$, waarbij:

- $f(i) \geq 0$, voor elke i
- en $\sum_{\text{alle } i} f(i) = 1$

De (kans)verdelingsfunctie van voorbeeld 1 is:

k	2	3	4	5	6	7	8	9	10	11	12
$F(k)$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{15}{36}$	$\frac{21}{36}$	$\frac{26}{36}$	$\frac{30}{36}$	$\frac{33}{36}$	$\frac{35}{36}$	$\frac{36}{36} = 1$

Voorbeeld 2

We gooien met een zuivere munt en definiëren K als het aantal pogingen totdat voor de eerste keer munt wordt gegooid. Dus $P(K = 1)$ betekent de kans dat bij de eerste worp

munt wordt gegooid. $P(K = 2)$ geeft de kans aan dat bij eerste worp kop is gegooid en bij de tweede worp munt, enzovoorts. Doordat kop en munt bij een zuiver muntstuk dezelfde kans hebben en de pogingen onafhankelijk zijn, geldt:
$P(K = 1) = \frac{1}{2}$, $P(K = 2) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2^2}$, $P(K = 3) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2^3}$, enzovoorts.
De kansfunctie voor dit experiment is: $f(k) = \dfrac{1}{2^k}$ $(k = 1, 2, 3, \ldots)$.
De waarde van bijvoorbeeld $F(3)$ is:
$F(3) = P(K \leq 3) = \sum\limits_{i=1}^{3} f(i) = \sum\limits_{i=1}^{3} \dfrac{1}{2^i} = \dfrac{1}{2^1} + \dfrac{1}{2^2} + \dfrac{1}{2^3} = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} = \frac{7}{8}$

Uit de definitie van de de verdelingsfunctie $F(k)$ kunnen de volgende eigenschappen worden afgeleid:

Eigenschap 1
De verdelingsfunctie is een niet-dalende functie. Daar $f(k) \geq 0$ voor elke k, zal bij het toenemen van k de verdelingsfunctie $F(k) = \sum\limits_{i \leq k} f(i)$ ook alleen maar toenemen of gelijk blijven (in elk geval niet afnemen).

Eigenschap 2
$\lim\limits_{k \to \infty} F(k) = 1$ (maar meestal is de grootst mogelijke waarde van k gelijk aan de steekproefgrootte n)

Eigenschap 3
$\lim\limits_{k \to -\infty} F(k) = 0$ (maar meestal is de kleinst mogelijke waarde van k gelijk aan 0)

5.3 Parameters van een (discrete) kansverdeling

Zoals in de inleiding van dit hoofdstuk is gezegd, kan een kansverdeling net als een frequentieverdeling worden gekarakteriseerd door parameters. De bekendste parameter is de verwachtingswaarde (ofwel het gemiddelde) van de kansverdeling.

5.3.1 Verwachtingswaarde van een discrete kansverdeling

In hoofdstuk 3 hebben we het begrip verwachtingswaarde al even toegelicht. In deze paragraaf zullen we wat dieper op de betekenis van *verwachtingswaarde* (kortweg *verwachting*) ingaan. Dit doen we aan de hand van een eenvoudig voorbeeld.

Voorbeeld 3
In een café wordt het volgende spel gespeeld. Een bargast zet een bepaald bedrag in. Een muntstuk wordt opgegooid en de caféhouder betaalt de bargast 10 eurocent als er kop wordt gegooid en 20 eurocent als er munt wordt gegooid. De caféhouder houdt het muntstuk om het spel steeds met hetzelfde muntstuk te kunnen spelen. Iedere keer is de

inzet dezelfde. Hoe hoog moet de inzet zijn zodat de caféhouder niet kan worden beticht van uitzuigerij, maar er ook geen reden is dat hij failliet gaat? Met andere woorden: wat moet de inzet zijn opdat er sprake is van een 'eerlijk' spel?

Oplossing

Als het experiment een groot aantal keren wordt gespeeld, zal op den duur in ongeveer de helft van het aantal worpen kop vallen en in de andere helft van de gevallen munt. Dit gaat op als het muntstuk zuiver is.
Wordt het spelletje N keer uitgevoerd, dan zal de caféhouder moeten uitkeren:
$\frac{N}{2} \times 0{,}10 + \frac{N}{2} \times 0{,}20 = \frac{N}{2} \times 0{,}30 = N \times 0{,}15$
Om over een lange periode precies quitte te spelen, zal de caféhouder een inzet van 0,15 eurocent moeten vragen. De gemiddeld te verwachten betaling op de lange duur is namelijk $\frac{1}{2} \cdot 0{,}10 + \frac{1}{2} \cdot 0{,}20 = 0{,}15$.

Het gemiddelde op de lange duur noemt men nu de *verwachtingswaarde* of *verwachting* (Engels: *mathematical expectation*, kortweg *expectation* met als schrijfwijze E).

Voorbeeld 4

Bij een bepaalde loterij is de kans op een prijs van 5000 euro gelijk aan 0,01%. De kans van op een prijs van 2000 euro is gelijk aan 0,05%. Bepaal de verwachtingswaarde van het te winnen bedrag.

Oplossing

De verwachtingswaarde van het te winnen bedrag is:
$E(\text{'winst'}) = 0{,}0001 \cdot 5000 + 0{,}0005 \cdot 2000 = 1{,}50$ euro.

We kunnen nu definiëren:

Definitie

Stel dat K een discrete kansvariabele is, die de waarden $k_1, k_2, ..., k_n$ kan aannemen met respectievelijk de kansen $p_1, p_2, ..., p_n$, waarbij $\sum_{i=1}^{n} p_i = 1$. Dan is de verwachtingswaarde van K gelijk aan $\sum_{i=1}^{n} p_i k_i$.

In formulevorm: $E(K) = \sum_{i=1}^{n} p_i k_i$ of (met $p_i = f(k_i)$, de kansfunctie)

$$E(K) = \sum_{i=1}^{n} k_i f(k_i) \qquad (5.1)$$

Soms schrijft men μ in plaats van $E(K)$. De verwachtingswaarde is te beschouwen als het *gewogen gemiddelde* van alle uitkomsten. Als *wegingsfactor* voor de uitkomst k_i in dit gewogen gemiddelde functioneert de kans(functie) $p_i = f(k_i)$.

Voorbeeld 5

Jan en Henk spelen een dobbelspel. Jan gooit met een dobbelsteen. Henk keert daarbij de volgende bedragen uit:
10 eurocent bij een 1 of een 2
20 eurocent bij een 3 of een 4
40 eurocent bij een 5
80 eurocent bij een 6
Hoe groot is de winst(verwachting) van Jan?

Oplossing

Als K de 'uitbetaling' is, dan is $P(K = 10) = \frac{2}{6}$, $P(K = 20) = \frac{2}{6}$, $P(K = 40) = \frac{1}{6}$ en $P(K = 80) = \frac{1}{6}$.
De winstverwachting van Jan is:
$E(K) = 10 \cdot \frac{2}{6} + 20 \cdot \frac{2}{6} + 40 \cdot \frac{1}{6} + 80 \cdot \frac{1}{6} = \frac{180}{6} = 30$ eurocent.

In voorbeeld 5 kunnen we de gebeurtenis K ook definiëren als '*Het ogental bij een worp met een dobbelsteen*'. Verder beschouwen we de kansvariabele L, welke de 'verdiensten' aangeeft van Jan. Deze verdiensten zijn een functie van K, aangegeven door $g(K)$, met de volgende kansverdeling:

K	1	2	3	4	5	6
$L = g(K)$	10	10	20	20	40	80

De verwachtingswaarde van K hangt af van de functie g van K; dit noteren we als $E\{g(K)\}$, waarbij geldt:
$E\{g(K)\} = \sum_{\text{alle } i} g(k_i) \cdot p_i$ of, anders geschreven, $E\{g(K)\} = \sum_{\text{alle } i} g(k_i) \cdot f(k_i)$, waarin f de kansfunctie is van K en $p_i = f(k_i)$.
Bij uitwerking ontstaat hetzelfde antwoord als hierboven gegeven.

Opdracht

Ga dit na, door de berekening voor voorbeeld 5 uit te voeren.

We kunnen nu ook de verwachting van $L^2 = \{g(K)\}^2$ bepalen. Volgens de gegeven definitie (formule (5.1)) is:

$$E\{g(K)\}^2 = \sum_{\text{alle } i} g(k_i)^2 \cdot p_i = \sum_{\text{alle } i} L_i^2 \cdot p_i$$

Uitgewerkt voor voorbeeld 5 krijgen we:

$L^2 = g(k)^2$	100	100	400	400	1600	6400
p_i	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Daaruit volgt:
$E\{g(K)\}^2 = 100 \cdot \frac{2}{6} + 400 \cdot \frac{2}{6} + 1600 \cdot \frac{1}{6} + 6400 \cdot \frac{1}{6} = \frac{9000}{6} = 1500$ (eurocent)^2.

Eigenschappen van de verwachtingswaarde van een discrete kansverdeling

We gaan een aantal belangrijke eigenschappen van de verwachtingswaarde van een discrete kansvariabele behandelen. De bewijzen van deze eigenschappen laten we achterwege.

Eigenschap 4

De verwachtingswaarde is een zogenaamde lineaire operator, dat wil zeggen als a en b constante getallen zijn, dan geldt:

1. $E(K + a) = E(K) + a$
2. $E(bK) = b \cdot E(K)$
3. $E(bK + a) = b \cdot E(K) + a$

Opmerking

De derde eigenschap is een samentrekking van de eerste twee.

Voorbeeld 6

Stel K kan de waarden 3, 4 en 5 aannemen met respectievelijk de kansen 0,1; 0,3 en 0,6. De verwachtingswaarde van $6K + 2$ verkrijgen we door eerst $E(K)$ te berekenen:
$E(K) = 0{,}1 \cdot 3 + 0{,}3 \cdot 4 + 0{,}6 \cdot 5 = 4{,}5$
$E(6K + 2) = 6E(K) + 2 = 6 \cdot 4{,}5 + 2 = 29$

Opdracht

Onderzoek het antwoord in voorbeeld 6 door de waarden 3, 4 en 5 te vermenigvuldigen met 6 en bij de verkregen uitkomsten 2 op te tellen. Bepaal van deze drie nieuwe waarden de verwachtingswaarde.

Eigenschap 5

Indien $g(K)$ een functie is van K, geldt:
$E\{b \cdot g(K) + a\} = b \cdot E\{g(K)\} + a$

Voorbeeld 7

Stel dat K de waarden 1, 2 en 3 kan aannemen. Stel verder dat de kans in alle drie de gevallen gelijk is aan $\frac{1}{3}$. De verwachtingswaarde van $3K^2 + 4$ wordt als volgt gevonden:
$E(3K^2 + 4) = 3E(K^2) + 4 = 3(1^3 \cdot \frac{1}{3} + 2^3 \cdot \frac{1}{3} + 3^3 \cdot \frac{1}{3}) + 4 = 3(\frac{1}{3} + \frac{8}{3} + \frac{28}{3}) + 4 = 40$

Eigenschap 6

Zijn K en L twee discrete kansvariabelen, dan geldt voor hun som Z:

$$E(Z) = E(K + L) = E(K) + E(L) \tag{5.2}$$

Voorbeeld 8

K kan de waarden 3 en 4 aannemen, met respectievelijk de kansen 0,4 en 0,6. L kan de waarden 3, 4 en 6 aannemen, met respectievelijk de kansen 0,45 en 0,3 en 0,25. De verwachtingswaarde van $Z = 3K + 2L$ is dan:
$E(Z) = E(3K + 2L) = E(3K) + E(2L) = 3E(K) + 2E(L)$
$E(K) = 0{,}4 \cdot 3 + 0{,}6 \cdot 4 = 3{,}6$
$E(L) = 0{,}45 \cdot 3 + 0{,}3 \cdot 4 + 0{,}25 \cdot 6 = 4{,}05$, zodat
$E(Z) = 3 \cdot 3{,}6 + 2 \cdot 4{,}05 = 18{,}9$

5.3.2 De variantie van een discrete kansvariabele

De mate van spreiding van de mogelijke waarden van een kansvariabele K ten opzichte van de verwachtingswaarde $E(K)$ wordt weergegeven door de *variantie*.
Voor de variantie van K worden de volgende schrijfwijzen gebruikt: $var(K)$, $\sigma^2(K)$ of σ_K^2.

Definitie

De variantie van K is de verwachtingswaarde van het kwadraat van de afwijkingen van K ten opzichte van $E(K)$. De definitie van de variantie kunnen we als volgt weergeven:

$$var(K) = E\{K - E(K)\}^2 \tag{5.3}$$

of na herleiding tot een andere schrijfwijze:

$$var(K) = E(K^2) - \{E(K)\}^2 \tag{5.3a}$$

Op de afleiding van de herleiding van formule (5.3) naar formule (5.3a) gaan we niet in. Wel zij opgemerkt dat deze eigenschap reeds genoemd is aan het eind van hoofdstuk 3 (zie formule 3.12).
Gezien de definitie van de verwachting is de variantie van een discrete kansvariabele volgens formule (5.3) een *gewogen gemiddelde* van de kwadratische afwijkingen van de mogelijke uitkomsten ten opzichte van de verwachtingswaarde. Als weegfactoren fungeren de kansen op de verschillende uitkomsten k_i.
$var(K) = \sum_{i=1}^{n} p_i\{(k_i - E(K)\}^2$ of (na herleiding) $var(K) = \sum_{i=1}^{n} p_i k_i^2 - \{E(K)\}^2$. Naar analogie van formule (5.1), waarin p_i vervangen is door de kansfunctie $f(k_i)$ kunnen we schrijven:

$$var(K) = \sum_{i=1}^{n} f(k_i)\{(k_i - E(K)\}^2 \quad (5.4)$$

$$= \sum_{i=1}^{n} f(k_i) \cdot k_i^2 - \{E(K)\}^2 \quad (5.4a)$$

Voorbeeld 9

De discrete kansvariabele K kan de waarden 4, 5 en 6 aannemen, met respectievelijk de kansen 0,4; 0,2 en 0,4. De variantie van K wordt als volgt berekend:

k_i	$f(k_i)$	$k_i \cdot f(k_i)$	$f(k_i) \cdot k_i^2$
4	0,4	1,6	6,4
5	0,2	1,0	5,0
6	0,4	2,4	14,4
$\sum$	1,0	5,0	25,8

$E(K) = \sum k_i \cdot f(k_i) = 5{,}0$ zodat volgens formule (5.4a)
$var(K) = 25{,}8 - (5{,}0)^2 = 25{,}8 - 25{,}0 = 0{,}8$

Opdracht

Controleer dit resultaat door ook formule (5.4) toe te passen.

Zoals bekend (zie hoofdstuk 3) kennen we naast de variantie nog een tweede spreidingsmaat, de standaardafwijking, die veelvuldig wordt gebruikt om de spreiding van een kansvariabele weer te geven.
De standaardafwijking is de wortel uit de variantie. In formulevorm geldt dus

$$\sigma_K = \sqrt{var(K)} = \sqrt{E(K^2) - \{E(K)\}^2}$$

Voorbeeld 10

De standaardafwijking van de kansvariabele in voorbeeld 9 is:
$\sigma_K = var(K) = \sqrt{0{,}8} \approx 0{,}9$

5.4 Theoretische discrete kansverdelingen

In de volgende subparagrafen bespreken we een aantal discrete kansverdelingen, die voor de praktijk erg belangrijk zijn. Daarbij onderscheiden wij:

- de *binomiale verdeling*
- de *hypergeometrische verdeling*
- de *Poisson-verdeling*

We zullen in de komende paragrafen deze verdelingen behandelen.

5.4.1 De binomiale verdeling

Ter inleiding van de *binomiale verdeling* doen we het volgende experiment:
we werpen met drie zuivere dobbelstenen en beschouwen de gebeurtenis: 'de dobbelsteen heeft vier ogen boven'.
Ten aanzien van een dobbelsteen kunnen we in dit geval twee kenmerken onderscheiden:
a. een dobbelsteen heeft vier ogen boven;
b. een dobbelsteen heeft geen vier ogen boven.
We definiëren nu: K is 'het aantal stenen met vier ogen boven'.
De kansvariabele K kan de waarden 0, 1, 2 of 3 aannemen, want met drie dobbelstenen kan men 0, 1, 2 of 3 'vieren' gooien.

$P(K = 3)$, dus de kans dat alle drie dobbelstenen een vier opleveren, is:

$$P(K = 3) = \tfrac{1}{6} \cdot \tfrac{1}{6} \cdot \tfrac{1}{6} = \left(\tfrac{1}{6}\right)^3$$

De kans op geen enkele vier, $P(K = 0)$ is op dezelfde manier te berekenen:

$$P(K = 0) = \tfrac{5}{6} \cdot \tfrac{5}{6} \cdot \tfrac{5}{6} = \left(1 - \tfrac{1}{6}\right)^3$$

Voor het berekenen van de kans $P(K = 1)$ is toepassing van alleen de vermenigvuldigingsregel voor onafhankelijke gebeurtenissen niet voldoende, omdat $K = 1$ op drie, elkaar uitsluitende manieren kan worden gerealiseerd, namelijk:

mogelijkheid	I	II	III	kans
A	4	$\bar{4}$	$\bar{4}$	$\frac{1}{6} \cdot \frac{5}{6} \cdot \frac{5}{6}$
B	$\bar{4}$	4	$\bar{4}$	$\frac{5}{6} \cdot \frac{1}{6} \cdot \frac{5}{6}$
C	$\bar{4}$	$\bar{4}$	4	$\frac{5}{6} \cdot \frac{5}{6} \cdot \frac{1}{6}$

(Een streepje boven de 4 wil zeggen: 'niet vier'.)
In hoofdstuk 4 hebben we geleerd dat het aantal combinaties van één vier en (dus) twee 'niet-vieren' gelijk is aan $\binom{3}{1}$ respectievelijk $\binom{3}{2}$. Anders gezegd: we kunnen op $\binom{3}{1} = \binom{3}{2} = 3$ manieren, met drie dobbelstenen, één vier gooien.
Daar elk van de drie elkaar uitsluitende mogelijkheden A, B en C tot de *gunstige* gevallen of *successen* behoort, dient de optelregel te worden toegepast. Er moeten drie dezelfde kansen worden opgeteld.

De kans op *één vier*, $P(K = 1)$ is als volgt te berekenen:

$$P(K=1) = \binom{3}{1}\left(\tfrac{1}{6}\right)^1\left(\tfrac{5}{6}\right)^2 = 3\cdot\left(\tfrac{1}{6}\right)^1\left(1-\tfrac{1}{6}\right)^2$$

Ook voor de uitkomst $K = 2$ bestaan drie mogelijkheden:

$(4, 4, \bar{4})$, $(4, \bar{4}, 4)$ en $(\bar{4}, 4, 4)$ elk met een kans $\left(\frac{1}{6}\right)^2 \cdot \frac{5}{6}$, dus:

$$P(K=2) = \binom{3}{2}\left(\tfrac{1}{6}\right)^2\left(\tfrac{5}{6}\right)^1 = 3\cdot\left(\tfrac{1}{6}\right)^2\left(1-\tfrac{1}{6}\right)^1$$

Resumerend in tabelvorm:

k	kans	anders geschreven	uitgerekend
0	$(1-\frac{1}{6})^3$	$\binom{3}{0}\left(\frac{1}{6}\right)^0\left(1-\frac{1}{6}\right)^{3-0}$	$=\frac{125}{216}$
1	$3\cdot\left(\frac{1}{6}\right)^1(1-\frac{1}{6})^2$	$\binom{3}{1}\left(\frac{1}{6}\right)^1\left(1-\frac{1}{6}\right)^{3-1}$	$=\frac{75}{216}$
2	$3\cdot\left(\frac{1}{6}\right)^2(1-\frac{1}{6})^1$	$\binom{3}{2}\left(\frac{1}{6}\right)^2\left(1-\frac{1}{6}\right)^{3-2}$	$=\frac{15}{216}$
3	$\left(\frac{1}{6}\right)^3$	$\binom{3}{3}\left(\frac{1}{6}\right)^3\left(1-\frac{1}{6}\right)^{3-3}$	$=\frac{5}{216}$
		$\sum$	$=\frac{216}{216}=1$

Als controlemiddel tellen we de kansen op en zien dan dat de som van alle kansen zoals verwacht 1 is.

De derde kolom levert de mogelijkheid om te komen tot een algemene schrijfwijze voor de berekening van kansen bij dit soort problemen. We voeren daarbij de volgende symbolen en omschrijvingen in, met tussen haakjes de situatie in ons experiment:

n = aantal elementen in een steekproef (aantal dobbelstenen $n = 3$);

p = kans op het optreden van een bepaalde gebeurtenis (kans op een vier bij een dobbelsteen: $p = \frac{1}{6}$);

k = aantal elementen in de steekproef waarbij die gebeurtenis optreedt (aantal stenen waarbij een vier bovenkomt).

De kans dat de discrete kansvariabele K een bepaalde waarde k aanneemt (de kansverdeling van K) kan nu algemeen worden geschreven als:

$$P(K=k) = \binom{n}{k}p^k(1-p)^{n-k}, \text{ met } k = 0, 1, ..., n \text{ en } 0 \le p \le 1 \qquad (5.5)$$

De kansvariabele K, met bovenstaande kansfunctie, heet een ***binomiale kansvariabele***.

Voorbeeld 11

We werpen vijf keer achter elkaar met een *onzuiver* muntstuk, waarbij de kans op 'kop' is p. De kans op munt is dan $1 - p$.

De kans op driemaal kop, dus tweemaal munt, kan als volgt berekend worden:

In een rij van 5 uitkomsten bestaande uit 3 maal kop en 2 maal munt zijn $\binom{5}{3}$ verschillende volgorden of permutaties mogelijk. De kans op het voorkomen van een rij bestaande uit 3 maal kop en 2 maal munt (maar in willekeurige volgorde!) is:

$$P(K = 3) = \binom{5}{3} p^3 (1 - p)^2$$

5.4.2 Voorwaarden voor toepassing van de binomiale verdeling

De binomiale verdeling kan gebruikt worden in de volgende omstandigheden.

- *Bij het nemen van een steekproef uit een eindige populatie.*

  Stel dat in een populatie twee kenmerken vertegenwoordigd zijn. Bij een steekproef mét teruglegging mag de binomiale verdeling worden toegepast. De kans p (ook wel *fractie* genoemd) om bij een trekking uit de populatie een bepaald kenmerk aan te treffen dat in de populatie aanwezig is, is dan namelijk voor iedere trekking hetzelfde. We hebben hierbij te maken met onafhankelijke gebeurtenissen.

  Bij steekproeven zónder teruglegging (dus bij afhankelijke gebeurtenissen) mag ook de binomiale verdeling worden toegepast, mits de populatieomvang N tenminste 10 maal de steekproefgrootte (n) is. Wanneer aan deze vuistregel voldaan is, blijft tijdens het nemen van de steekproef de verhouding van het aantal elementen met een bepaald kenmerk en het aantal elementen zonder dat kenmerk bij benadering constant. Hoewel we in principe met afhankelijkheid te maken hebben, mogen we (mits aan de vuistregel $N \geq 10n$ voldaan is) doen alsof we te maken hebben met onafhankelijkheid.

- *Bij het nemen van steekproeven uit een oneindig grote populatie.*

  De kans p op het aantreffen van een bepaald kenmerk is dan sowieso constant. Wanneer bijvoorbeeld de voetbaltoto op willekeurige wijze wordt ingevuld, is de kans op een goede voorspelling voor iedere wedstrijd gelijk aan $\frac{1}{3}$. Het aantal goed voorspelde wedstrijden is dus te beschouwen als een steekproef uit een oneindig grote populatie met teruglegging. Het aantal 'successen' (goed gegokte uitslagen) is binomiaal verdeeld.

5.4.3 Verwachtingswaarde en variantie van de binomiale verdeling

De kansfunctie van de binomiale verdeling $f(k) = P(K = k) = \binom{n}{k} p^k (1-p)^{n-k}$ voldoet aan de definitie van een kansfunctie. Voor elke waarde van k ($k = 0, 1, 2, 3, ..., n$) geldt dat $P(K = k) \geqq 0$ en te bewijzen is dat $\sum_{k=0}^{n} P(K = k) = 1$.

Wanneer een discrete kansvariabele K een op deze kansfunctie gebaseerde verdeling bezit, zegt men dat K een binomiale verdeling bezit (of dat K binomiaal verdeeld is), met de parameters n en p.

Voor het *gemiddelde* of *verwachtingswaarde* $\mu = E(K)$ van een binomiale verdeling, met de parameters n en p geldt:

$$\mu = n \cdot p \tag{5.6}$$

Voor de *variantie* $\sigma^2 = var(K)$, respectievelijk de *standaardafwijking* $\sigma = \sqrt{var(K)}$ gelden:

$$\sigma^2 = n \cdot p \cdot (1 - p) \tag{5.7}$$

$$\sigma = \sqrt{n \cdot p \cdot (1 - p)} \tag{5.8}$$

Voorbeeld 12

Een Multiple-choicetoets bestaat uit 20 vragen met elk vier mogelijke antwoorden, waarvan er één goed is. Iemand besluit door willekeurig aan te kruisen (gokken) deze toets te maken. Bepaal de verwachtingswaarde en de standaardafwijking van het aantal goed gegokte vragen (=successen).

Oplossing

Het aantal 'successen' K is binomiaal verdeeld omdat $p = \frac{1}{4}$ constant is voor elke vraag. Aangezien $n = 20$ geldt: $E(K) = np = 20 \cdot \frac{1}{4} = 5$ en $\sigma = \sqrt{n \cdot p \cdot (1 - p)} = \sqrt{20 \cdot \frac{1}{4} \cdot \frac{3}{4}}$. De standaardafwijking is dus $\sqrt{\frac{15}{4}} \approx 1{,}94$.

5.4.4 De tabel van de binomiale verdeling

De kansen van een binomiaal verdeelde kansvariabele K kunnen worden berekend met de *binomiaalformule*:

$$P(K = k) = \binom{n}{k} p^k (1 - p)^{n-k} \tag{5.9}$$

De binomiaalformule (formule (5.9)) is met een zakrekenmachine te berekenen, maar gemakkelijker gaat het met een programma als EXCEL (zie Appendix A). Voor het gemak is echter achter in dit boek ook een tabel (tabel B2) opgenomen met deze verdeling, althans voor $n \leq 10$, en enkele veel voorkomende waarden van de *fractie* p. Wanneer $n > 10$ kan de tabel van de binomiale verdeling niet worden gebruikt. In het volgende hoofdstuk zullen we zien dan in dat geval gebruik kan worden gemaakt van een benadering door middel van de normale verdeling. In de tabel zijn uitsluitend kansen opgenomen voor fracties $p \leq \frac{1}{2}$.

Opdracht

Bedenk waarom slechts waarden voor $p \leq \frac{1}{2}$ getabelleerd zijn, terwijl deze tabel toch ook bruikbaar is voor een aantal waarden voor $p > \frac{1}{2}$. Bereken vervolgens met behulp van tabel B2: $P(K = 3)$ en $P(K \leq 3)$ bij een steekproef van 8 stuks uit een populatie met fractie $p = 0{,}7$.

We zullen nu een aantal voorbeelden geven waarin de binomiale verdeling wordt toegepast.

Voorbeeld 13

De kans dat een geboren kind een meisje is, bedraagt 0,49. Bereken de kansverdeling van het aantal meisjes in een gezin met twee kinderen.

Oplossing

Omdat de fractie $p = 0{,}49$ constant is, is het aantal meisjes in een gezin van twee ($n = 2$) kinderen binomiaal verdeeld. Tabel B2 kan niet worden toegepast.
De kans dat onder twee aselect (willekeurig) gekozen geboorten 0, 1 of 2 meisjes zijn, is respectievelijk:

k	$P(K = k)$		
0	$\binom{2}{0} \cdot 0{,}49^0 \cdot 0{,}51^2 = 0{,}51^2$	=	0,2601
2	$\binom{2}{1} \cdot 0{,}49^1 \cdot 0{,}51^1 = 2 \cdot 0{,}49 \cdot 0{,}51^1$	=	0,4998
3	$\binom{2}{2} \cdot 0{,}49^2 \cdot 0{,}51^0 = 0{,}49^2$	=	0,2401
	$\sum$	=	1,0000

Voorbeeld 14

Van een beroepsziekte is bekend dat 10% van de mensen die dit beroep uitoefenen er door wordt aangetast. Wat is de kans dat van 3 nieuw aangenomen personeelsleden in dit beroep er hoogstens één de ziekte krijgt?

Oplossing

Het aantal personen dat de beroepsziekte krijgt, is binomiaal verdeeld met $p = 0{,}1$ en $n = 3$. Tabel B2 kan gebruikt worden:
Hoogstens één betekent $K = 0$ of $K = 1$.
De kans hierop is: $P(K \leq 1) = P(K = 0) + P(K = 1)$

k	$P(K = k)$
0	0,7290
1	0,2430
$\sum$	0,9720

De kans dat van 3 nieuwe personeelsleden er hoogstens één de beroepsziekte krijgt is dus: $P(K \leq 1) = 0{,}9720$.

Voorbeeld 15

Van een machine is bekend dat gemiddeld 8% van de geproduceerde exemplaren gedegradeerd wordt tot tweede keus. Uit de productie wordt nu een aselecte steekproef van $n = 20$ genomen. Wat is de kans dat er in de steekproef meer dan één fout exemplaar zit?

Oplossing

Het aantal 'tweede-keus' exemplaren is binomiaal verdeeld met $p = 0{,}08$. Tabel B2 kan niet toegepast worden.
De kans op meer dan één foutief exemplaar $P(K \geq 2) = 1 - P(K \leq 1)$
We hebben nu de volgende parameters: $n = 20$, $p = 0{,}08$.

k	$P(K = k)$		
0	$\binom{20}{0} \cdot 0{,}08^0 \cdot 0{,}92^{20}$	=	0,1887
1	$\binom{20}{1} \cdot 0{,}08^1 \cdot 0{,}92^{19}$	=	0,3282
	$\sum$	=	0,5169

De kans op meer dan één foutief exemplaar in een steekproef van 20 is dus
$P(K > 1) = 1 - P(K \leq 1) = 1 - 0{,}5169 = 0{,}4831$

Voorbeeld 16

Wat is de kans op 5 foutieve bouten in een aselecte steekproef van 10 bouten, als de kans op een foutieve bout 0,20 is?

Oplossing

We nemen aan dat de populatie waaruit de steekproef afkomstig is groter is dan 100 (10 keer de steekproefgrootte). Als K het aantal foutieve bouten voorstelt in een steekproef van 10, is K binomiaal verdeeld met $n = 10$ en $p = 0{,}2$. Tabel B2 kan toegepast worden:
Er geldt: $P(K = 5) = 0{,}0264$.

Voorbeeld 17

In een zeer grote loterij (veel loten) is de prijzenpot zodanig samengesteld dat 30% van de loten 'prijsloten' zijn. Iemand koopt 20 loten. Wat is de kans dat hij minstens 3 prijzen krijgt?

Oplossing

Als K het aantal prijsloten voorstelt in een partij van 20 loten, is K binomiaal verdeeld met $n = 20$ en $p = 0{,}3$. Tabel B2 kan toegepast worden.

Oplossing voorbeeld 17 (vervolg)

$P(K \geq 3) = 1 - \{P(K = 0) + P(K = 1) + P(K = 2)\} = 1 - P(K \leq 2)$

$P(K = 0) = 0{,}0008$
$P(K = 1) = 0{,}0068$
$P(K = 2) = 0{,}0278$
$P(K \leq 2) = 0{,}0354$

De kans op minstens 3 prijzen is daarom:
$P(K \geq 3) = 1 - P(K \leq 2) = 1 - 0{,}0354 = 0{,}9646.$

5.4.5 De hypergeometrische verdeling

We hebben gezien dat de binomiale verdeling geldt wanneer er sprake is van een steekproef met teruglegging uit een al dan niet eindige populatie. Daarnaast kan de binomiale verdeling worden toegepast bij een steekproef zonder teruglegging uit een populatie die minstens 10 maal groter is dan de steekproefgrootte. De binomiale verdeling mag dus *niet* worden gebruikt in de situatie dat er sprake is van een steekproef *zonder* teruglegging uit een populatie die *kleiner* is dan 10 maal de steekproefomvang. In dit geval is de parameter p niet constant voor elke gebeurtenis. De kansverdeling die nu in beeld komt is de *hypergeometrische kansverdeling*. Bij de hypergeometrische kansverdeling gaat men uit van een *eindige* populatie, waaruit een steekproef wordt getrokken *zonder* teruglegging. Dit betekent dat de populatie na trekking steeds kleiner wordt en de kans op een bepaald element met een zeker kenmerk, van trekking tot trekking verandert. De fractie p is niet constant.

De hypergeometrische formule

Stel dat uit een (kleine) partij van N geproduceerde artikelen, waarvan er M defect zijn, een aselecte steekproef zonder teruglegging wordt getrokken van n artikelen. Gevraagd: de kans dat zich in de steekproef precies k defecte (dus $n - k$ niet-defecte) artikelen bevinden. Deze kans kunnen we het gemakkelijkst berekenen door gebruik te maken van de in hoofdstuk 4 aangeleerde begrippen uit de combinatoriek.
Aan het aantal 'defecte artikelen' kennen we de kansvariabele K toe. De kansvariabele K is een hypergeometrisch verdeelde kansvariabele.
De kansfunctie van K is te geven door de *hypergeometrische formule*:

$$f(k) = P(K = k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} \text{ voor } k = 0, 1, 2, \ldots, n \tag{5.10}$$

Formule (5.10) kunnen we als volgt interpreteren:

- N is het aantal elementen (artikelen) in de partij (=populatie);
- M is het aantal elementen (artikelen) in de populatie met een bepaalde eigenschap ('defect');
- n is het aantal elementen (artikelen) in de steekproef;
- k is het aantal elementen (artkelen) in de steekproef met een bepaald eigenschap ('defect').

  In de formule van de kansfunctie kan men drie verschillende factoren beschouwen, die alle drie gebaseerd zijn op *het aantal combinaties.* De volgorde is immers niet van belang (de artikelen zijn niet eens genummerd).
- $\binom{M}{k}$ is het aantal verschillende manieren om uit M elementen, die een bepaalde eigenschap (defect) bezitten, precies k elementen te trekken. Bij al deze manieren zijn er $\binom{N-M}{n-k}$ verschillende manieren om uit $N - M$ elementen, die een bepaalde eigenschap *niet* (dus hier: *niet* defect, oftewel kwalitatief goed) bezitten, precies $n - k$ elementen te trekken. Het aantal gunstige manieren om k elementen met het kenmerk (hier: defect) en $n - k$ elementen zonder dat kenmerk (niet defect) te trekken is dan $\binom{M}{k}\binom{N-M}{n-k}$.

Het aantal mogelijke manieren om een steekproef van n stuks uit een partij van N stuks te trekken, is $\binom{N}{n}$. Door nu de klassieke kansdefinitie $\left(\text{kans} = \frac{\text{aantal gunstige manieren}}{\text{aantal mogelijke manieren}}\right)$ toe te passen ontstaat formule (5.10).

Voorbeeld 18

Om een prijs te winnen in een lotto moet men een aantal genummerde balletjes goed hebben van de zes balletjes die zonder teruglegging zijn getrokken uit een populatie van balletjes die genummerd zijn van 1 tot 45. Wat is de kans op vier goede nummers?

Oplossing

$$P(K = 4) = f(4) = \frac{\binom{6}{4}\binom{45-6}{6-4}}{\binom{45}{6}} = \frac{\binom{6}{4}\binom{39}{2}}{\binom{45}{6}} = \frac{\frac{6!}{4!2!}\cdot\frac{39!}{2!37!}}{\frac{45!}{6!39!}} = \frac{741}{543004}$$

$$= 1,3646 \cdot 10^{-3}$$

Opdracht

Wat is de kans op 6 goede nummers?

Voorbeeld 19

In een partij van 40 computers zitten 4 defecte computers. Men trekt uit de partij aselect zonder terugzetting 2 computers. Bepaal de kansverdeling van het aantal defecte computers in de steekproef.

Oplossing

Met behulp van de hypergeometrische kansverdeling kan men de verschillende kansen bepalen. We hebben de volgende parameters: $N = 40$, $M = 4$, $n = 2$.

$$f(0) = P(K = 0) = \frac{\binom{4}{0}\binom{40-4}{2-0}}{\binom{40}{2}} = \frac{\binom{4}{0}\binom{36}{2}}{\binom{40}{2}} = \frac{\frac{4!}{0!4!}\cdot\frac{36!}{2!34!}}{\frac{40!}{2!38!}} = \tfrac{630}{780} = 0{,}8077$$

$$f(1) = P(K = 1) = \frac{\binom{4}{1}\binom{40-4}{2-1}}{\binom{40}{2}} = \frac{\binom{4}{1}\binom{36}{1}}{\binom{40}{2}} = \frac{\frac{4!}{1!3!}\cdot\frac{36!}{1!35!}}{\frac{40!}{2!38!}} = \tfrac{144}{780} = 0{,}1846$$

$$f(2) = P(K = 2) = \frac{\binom{4}{2}\binom{40-4}{2-2}}{\binom{40}{2}} = \frac{\binom{4}{2}\binom{36}{0}}{\binom{40}{2}} = \frac{\frac{4!}{2!2!}\cdot\frac{36!}{0!36!}}{\frac{40!}{2!38!}} = \tfrac{6}{780} = 0{,}0077$$

De kansverdeling ziet er nu als volgt uit:

K	$f(k)$
0	0,8077
1	0,1846
2	0,0077
$\sum$	1,0000

We zien dat de som van alle kansen gelijk is aan 1. We hebben dus inderdaad te doen met een kansverdeling.

Opmerking

We hadden de kansen uit bovenstaand voorbeeld ook direct met de rekenregels uit de kansrekening kunnen berekenen in plaats van met de hypergeometrische formule. We geven daarbij een goede computer in de steekproef aan met G en een defecte computer aan met D. De kans op respectievelijk 0, 1 en 2 defecte computers in de steekproef van 2 computers uit een partij van 40 computers is:

- $f(0) = P(GG) = \frac{36}{40} \cdot \frac{35}{39} = 0{,}8077$
- $f(1) = P(DG) + P(GD) = \frac{4}{40} \cdot \frac{36}{39} + \frac{36}{40} \cdot \frac{4}{39} = 0{,}1846$
  (let op: er zijn twee elkaar uitsluitende mogelijkheden)
- $f(2) = P(DD) = \frac{4}{40} \cdot \frac{3}{39} = 0{,}0077$

We zien dat we dezelfde kansverdeling hebben gevonden als door middel van de hypergeometrische verdeling. Soms is het gemakkelijker direct via de kansregels tot een oplossing te komen, een andere keer geeft de hypergeometrische verdeling een snellere oplossing. Vooral bij wat complexere vraagstukken maken we liever gebruik van de hypergeometrische verdeling.

We kunnen de formule voor de hypergeometrische verdeling verder uitbreiden.

Voorbeeld 20

Op een schaal liggen 10 gebakjes (5 vruchtengebakjes, 3 slagroompunten en 2 tompouces). Iedere gast kiest aselect een gebakje. Hoe groot is de kans dat de zesde gast alleen nog kan kiezen uit 1 slagroompunt en 4 vruchtengebakjes?

Oplossing

De kans dat de eerste vijf gasten, 1 vruchtengebakje uit 5 mogelijkheden, 2 slagroompunten uit 3 mogelijkheden en 2 tompouces uit 2 mogelijkheden hebben genomen is, als wij de onafhankelijkheid van de keuze van de gasten veronderstellen:

$$P(1V, 2S, 2M) = \frac{\binom{5}{1}\binom{3}{2}\binom{2}{2}}{\binom{10}{5}} = \frac{\frac{5!}{1!4!} \cdot \frac{3!}{2!1!} \cdot \frac{2!}{2!0!}}{\frac{10!}{5!5!}} = \frac{5 \cdot 3 \cdot 1}{252} = \frac{15}{252} = 0{,}0595$$

Indien in de hypergeometrische verdeling n veel kleiner is dan N, is te verwachten dat het weinig verschil maakt of de trekkingen met of zonder teruglegging geschieden. We kunnen dan de binomiale verdeling toepassen.

Opdracht

Bereken de kansen uit voorbeeld 19 nogmaals door de binomiale verdeling toe te passen (dit mag want $n \leq 0{,}1N$) en merk op hoe klein het verschil is.

5.4.6 De Poisson-verdeling

We zullen nu een discrete kansverdeling bespreken die geïntroduceerd is in 1837 door S.D. Poisson. Deze verdeling kunnen we, wiskundig gezien, op twee manieren afleiden. Ten eerste geheel 'zelfstandig', maar we kunnen de Poisson-verdeling ook beschouwen als een *limietgeval* van de binomiale verdeling.
Om het verband met de binominale verdeling duidelijk te laten uitkomen, zullen wij de Poisson-verdeling afleiden als een limiet van de binomiale verdeling.
In de praktijk zijn er talloze situaties waarbij aan alle voorwaarden voor een binominale verdeling lijkt te zijn voldaan. Bij de uitwerking en interpretatie van deze gevallen stuit men echter op grote moeilijkheden. We zullen dit toelichten met een voorbeeld.

Voorbeeld 21

Aan een loket melden zich gemiddeld 60 personen per uur. Hoe is het aantal klanten dat zich per uur aan het loket meldt verdeeld?

Oplossing

Ideaal zou zijn als er precies elke minuut iemand bij het loket zou aankomen en binnen een minuut weer zou vertrekken. In dat geval is het aantal klanten dat zich per uur aan het loket meldt altijd 60, dus constant. In de praktijk zal, als we de aankomsten per minuut bekijken, wel gemiddeld één persoon per minuut arriveren, maar er zullen veel minuten zijn waarin niemand binnenkomt en ook minuten, waarin er 1, 2, 3 of meer aankomsten zijn.
Gesteld dat iemand precies één seconde nodig heeft om binnen te komen, dan is de kans dat in een bepaalde seconde iemand arriveert: $\frac{60}{3600} = \frac{1}{60}$ en de kans dat niet iemand aankomt $\frac{59}{60}$. Het aantal klanten dat per minuut aankomt lijkt dus binomiaal verdeeld met fractie $\frac{1}{60}$. In een uur zijn er 3600 momenten waarop iemand binnen kan komen, maar het aantal momenten waarop dit werkelijk gebeurt, is in verhouding tot het aantal mogelijke keren erg klein.
Kiest men een nog kleinere tijdseenheid (bijvoorbeeld 0,001 seconde), dan is de kans op het optreden van een gebeurtenis (aankomst) binnen die tijdseenheid vrijwel nul en het aantal mogelijke keren (binnen een tijdsbestek van een uur) wordt 1000 keer groter. Waar blijven we nu met de binomiale verdeling?
We zullen deze vraag hierna beantwoorden.

In voorbeeld 21 zien we dat het aantal mogelijke gebeurtenissen (aankomsten), per tijdseenheid zeer groot is, maar de kans op zo'n gebeurtenis is erg klein.
Het totaal aantal mogelijke gebeurtenissen (n) is zeer groot en nadert naar oneindig, terwijl p zeer klein is en naar nul nadert. In dit soort situaties maken we gebruik van de zogenaamde *Poisson-verdeling*.

5.4.7 Opbouw van een Poisson-verdeling

Het voorgaande kunnen we formeel samenvatten.

Heeft men een binominale kansvariabele met de parameters n en p, waarbij n nadert naar oneindig ($n \to \infty$) en p nadert naar nul ($p \to 0$), zodanig dat $n \cdot p$ constant blijft, dan kan dit limietgeval van de binominale verdeling benaderd worden door een Poisson-verdeling.

In het voorbeeld beschouwen we zeer kleine tijdsintervallen, waardoor het aantal momenten waarop iemand kan arriveren (n) nadert naar oneindig. Gelijktijdig zal de kans p, dat op een bepaald moment iemand arriveert, naar nul naderen. Als de aankomsten volkomen toevallig zijn en onderling onafhankelijk, zal voor $P(K = k)$ (de kans op k aankomsten) gelden:

$$P(K = k) = \lim_{\substack{n \to \infty,\ p \to 0 \\ \lambda = np \text{ is constant}}} \binom{n}{k} p^k (1 - p)^{n-k}$$

Bij het uitwerken van deze limiet (op de afleiding gaan we verder niet in) ontstaat de kansfunctie voor de variabele K, de *Poisson-formule*:

$$f(k) = P(K = k) = e^{-\lambda} \cdot \frac{\lambda^k}{k!} \text{ (met } k = 0, 1, 2, ..., n) \qquad (5.11)$$

In formule (5.11) is:

k = aantal 'successen'

λ = gemiddeld aantal 'successen' = $n \cdot p$

e = grondtal van de natuurlijke logaritme, waarbij e numeriek ongeveer gelijk is aan 2,71828...

Uit de formule voor de kansfunctie blijkt dat de verdeling van de Poisson-variabele geheel bepaald wordt door slechts één parameter, namelijk: λ.

Dat de Poisson-variabele een kansverdeling volgt, blijkt uit het feit dat men kan bewijzen dat de som van alle mogelijke kansen gelijk is aan 1, dus dat: $\sum_{k=0}^{n} P(K \le k) = \sum_{k=0}^{n} e^{-\lambda} \cdot \frac{\lambda^k}{k!} = 1$.

Het zijn met name aantallen gebeurtenissen per eenheid van tijd (het aantal verkeersongevallen per jaar, het aantal brandmeldingen per maand, het aantal *service-calls* per week, het aantal telefoongesprekken per dag, het aantal storingen per uur, het aantal bestellingen per kwartier, het aantal geigerteller-tikken per minuut) die een Poisson-verdeling bezitten. Maar ook voor andere dimensies zoals bijvoorbeeld lengte, oppervlakte en inhoud treffen we in de praktijk vaak Poisson-verdelingen aan, zoals bijvoorbeeld het aantal weeffouten per meter in een rol gordijnstof, het aantal oppervlaktefouten per cm^2 op de carosserie van een auto of het aantal bacteriën per cm^3 in de ons omringende lucht. Uit de bovenstaande voorbeelden blijkt dat we bij Poisson-verdeelde variabelen niet altijd kunnen spreken van zowel het aantal wel-optredende gebeurtenissen ('successen') als van het aantal niet-optredende gebeurtenissen ('mislukkingen'). Bij binomiale verdelingen kan dit juist wel.

5.4.8 De tabellen van de Poisson-verdeling

Voor verschillende waarden van λ is de Poisson-verdeling met de parameter λ vastgelegd in de tabellen B3 en B4. Tabel B3, de tabel van de *enkelvoudige Poisson-verdeling*, geeft voor een aantal opeenvolgende relevante waarden van K de kans $P(K = k)$ (formule (5.11)) en tabel B4, de tabel van de *cumulatieve Poisson-verdeling*, geeft voor een aantal opeenvolgende relevante waarden van c de kans $P(K \leq c) = P(K = 0) + P(K = 1) + P(K = 2) + ... + P(K = c)$.

Opdracht

Toon aan met behulp van tabel B3 dat in een Poisson-verdeling met $\lambda = 5$ voor $P(K \leq 4)$ dezelfde uitkomst verschijnt als in tabel B4 voor $c = 4$.

Daarnaast is de Poisson-formule uiteraard ook met een rekenmachine te berekenen.
In EXCEL is de Poisson-formule voorgeprogrammeerd.

Voorbeeld 22

Op een bepaald telefoontoestel komen gemiddeld 4 gesprekken per uur door. Het aantal gesprekken per uur is Poisson-verdeeld.

a. Wat is de kans dat er per uur precies twee gesprekken doorkomen?

b. En wat is de kans dat er per uur minstens twee gesprekken binnenkomen?

Oplossing

a. Stellen we het aantal telefoongesprekken dat per uur doorkomt K, dan volgt K een Poisson-verdeling met $\lambda = 4$. De kans dat er in een bepaald uur 2 gesprekken doorkomen, is: $P(K = 2) = e^{-\lambda} \cdot \frac{\lambda^k}{k!} = e^{-4} \cdot \frac{4^2}{2!} = e^{-4} \cdot \frac{16}{2} = 0{,}146$ (zie ook tabel B3).

b. De kans op tenminste 2 gesprekken in bovenstaand voorbeeld vinden we als volgt: $P(K \geqq 2) = 1 - P(K \leq 1) = 1 - 0{,}092 = 0{,}908$ oftewel 90,8%.
De kans $P(K \leq 1)$ is bepaald met tabel B4.

We moeten ons bij dit soort voorbeelden goed realiseren hoe de vraagstelling luidt. We kunnen bijvoorbeeld de volgende mogelijkheden onderscheiden:

- $P(\text{minder dan twee}) = P(K = 0) + P(K = 1) = P(K \leq 1)$
- $P(\text{ten hoogste twee}) = P(K = 0) + P(K = 1) + P(K = 2) = P(K \leq 2)$
- $P(\text{precies twee}) = P(K = 2)$
- $P(\text{minstens twee}) = P(K \geq 2) = 1 - P(K \leq 1)$
- $P(\text{meer dan twee}) = P(K \geq 3) = 1 - P(K \leq 2)$

In de volgende twee voorbeelden laten we toepassingen uit de praktijk zien.

Voorbeeld 23

Op een bepaald verkeersknooppunt in een stad gebeuren gemiddeld 0,8 aanrijdingen per dag (Poisson-verdeeld). Wat is de kans op maximaal drie aanrijdingen op één dag?

Oplossing

Het aantal aanrijdingen K is Poisson-verdeeld met $\lambda = 0{,}8$
De gevraagde kans is: $P(K \leq 3)$.
$P(K \leq 3) = 0{,}991$ (tabel B4).
De kans op maximaal 3 aanrijdingen bedraagt 0,991 of 99,1%

Voorbeeld 24

Op een garenspoel zit ongeveer 1000 m garen. Bij het opwikkelen op de spoel kunnen draadbreuken optreden. Van een bepaald garen is bekend dat er gemiddeld 1,2 draadbreuken per spoel optreden (Poisson-verdeeld). Wat is de kans op een spoel zonder draadbreuken?

Oplossing

$P(K = 0) = e^{-1{,}2} \cdot \frac{1{,}2^0}{0!} = 0{,}3012 \cdot 1 = 0{,}3012$ (zie ook tabel B3).
Conclusie: De kans op een draadbreukloze spoel is 30,12%.

5.4.9 Verwachtingswaarde en variantie van de Poisson-verdeling

In de inleiding hebben we de Poisson-verdeling voorgesteld als een limietgeval van de binomiale verdeling. De verwachtingswaarde van een binomiale verdeling is $E(K) = n \cdot p$. Voor de variantie geldt $var(K) = np(1 - p)$.
Op basis hiervan kunnen we vermoeden dat de verwachtingswaarde voor een Poisson (-verdeelde) variabele K gelijk zal zijn aan $E(K) = np = \lambda$. Dit is ook te bewijzen met behulp van formule (5.1) met formule (5.11) daarin ingevuld, maar we zullen dat bewijs niet geven.

Voor de variantie van een Poisson-variabele K geldt verrassenderwijs: $var(K) = np$. Dit is te bewijzen met formule (5.3) (en formule (5.11)). Bij nader inzien is dit resultaat ook weer niet zo verbazingwekkend. Kijk maar wat er gebeurt als in de formule $var(K) = np(1 - p)$ in de limietsituatie p naar 0 nadert, dus $1 - p$ naar 1.
Samenvattend zien we dat de verwachting, de variantie en de standaardafwijking van een Poisson-variabele gelijk zijn aan respectievelijk:

$$E(K) = \lambda \tag{5.12}$$

$$var(K) = \lambda \tag{5.13}$$

$$\sigma_K = \sqrt{\lambda} \tag{5.14}$$

Opmerking

Omdat de parameter λ van een Poisson-verdeling blijkbaar overeenkomt met de verwachtingswaarde (welke – zoals we weten – vaak met μ wordt aangeduid), wordt λ vaak direct vervangen door μ.

5.4.10 Optelbaarheid van Poisson-verdelingen

De Poisson-verdeling bezit een opmerkelijke eigenschap. Deze manifesteert zich wanneer twee Poisson-verdeelde variabelen worden opgeteld.

Stelling 1

Heeft men twee onderling onafhankelijke Poisson-variabelen K en L met de parameters λ_K en λ_L dan is de somvariabele $Z = K + L$ weer een Poisson-verdeling met de parameter $\lambda_Z = \lambda_K + \lambda_L$.

In de praktijk is dit een belangrijke regel. Stel dat bijvoorbeeld een product meerdere behandelingen ondergaat en elke behandeling heeft een kans op fouten, waarvan het aantal (per tijdseenheid of per serie) steeds Poisson-verdeeld is. Dan is volgens de stelling het totaal aantal fouten in het eindproduct ook weer Poisson-verdeeld. Nu volgt een ander voorbeeld.

Voorbeeld 25

Het aantal verkeersongelukken per maand met dodelijke afloop in plaats A is Poisson-verdeeld met parameter (dus verwachtingswaarde) 2. Dan is het aantal verkeersongelukken per jaar in plaats A eveneens Poisson-verdeeld, maar dan met parameter (gemiddelde) $12 \times 2 = 24$.

5.4.11 Benadering van een binomiale verdeling door een Poisson-verdeling

We hebben gezien dat de binomiale verdeling van de twee parameters n en p afhankelijk is. Zo'n verdeling is moeilijk te tabelleren. In het inleidend voorbeeld waarmee we deze paragraaf begonnen, is gesteld dat de Poisson-verdeling kan worden opgevat als een limietgeval van de binomiale verdeling. We kunnen nu in die gevallen, waarin n voldoende groot is en p voldoende klein, de binomiale verdeling benaderen door een Poisson-verdeling. De benadering past des te beter naarmate de waarde van n groter en de waarde van p kleiner wordt.

Als vuistregel kunnen we hanteren dat als $p < 0{,}1$ en $n > 25$, het verantwoord is de Poisson-verdeling te gebruiken als benadering voor een binomiale verdeling.

Voorbeeld 26

Een fabrikant keurt elke partij binnenkomende goederen door middel van een steekproef van 100 stuks. Wanneer in deze steekproef meer dan 5 foutieve exemplaren worden aangetroffen, wordt de betreffende partij afgekeurd en teruggezonden.

Wat is de kans dat een partij met 10% uitval bij deze controle zal worden afgekeurd?

Oplossing

In dit voorbeeld is $p = 0{,}1$ en $n = 100$. We kunnen dit binomiale probleem rekentechnisch benaderen door een Poisson-verdeling met de parameter $\lambda = np = 0{,}1 \cdot 100 = 10$.

Meer dan 5 fouten betekent $K \geq 6$.

$P(K \geq 6) = 1 - P(K \leq 5) = 1 - 0{,}067 = 0{,}933$ (tabel B4 is gebruikt).

De kans dat de partij met 10% uitval zal worden afgekeurd is dus 93,3%.

Opgaven

1. In de stad Dobbeldam is tussen 14.00 uur en 15.00 uur gemiddeld 1 van de 4 telefoonnummers in gesprek. Wanneer iemand in die tijd 4 keer opbelt, hoe groot is de kans, dat minstens één nummer in gesprek is?

2. In het station van Spoorstad komen gemiddeld 4 van de 5 treinen zonder vertraging binnen. Wanneer er per dag 6 treinen binnenkomen, hoe groot is de kans dat juist één trein met vertraging binnenkomt?

3. Van een beroepsziekte is bekend dat 25% van de mensen die dit beroep uitoefenen, er door wordt aangetast. Hoe groot is de kans dat van 7 nieuw aangenomen personeelsleden in dit beroep er hoogstens één van hen de beroepsziekte krijgt?

4. Een fabriek van huishoudelijke apparaten keurt elke partij binnengekomen onderdelen door middel van een steekproef van 20 stuks. Wanneer in deze steekproef meer dan 2 foutieve onderdelen worden gevonden, wordt betreffende partij afgekeurd. Bereken de kans dat een partij met 10% foutieve onderdelen wordt afgekeurd.

5. Een kwaliteitscontroleur, werkzaam bij een fabrikant van elektronicacomponenten, wil nagaan of een partij van een bepaalde component aan de specificatie voldoet (95% van de componenten werken goed). Hij neemt aselect een steekproef van 15 componenten uit de (grote) partij, die klaar staat voor verzending. De partij wordt goedgekeurd als alle 15 componenten goed functioneren.

a. Wat is de kans dat hij op basis van de steekproef de partij blokkeert (voldoet niet aan de specificatie), terwijl toch 95% van de componenten in de partij goed functioneert?

b. Wat is de kans dat hij de partij goedkeurt, terwijl slechts 90% van de componenten in de partij goed functioneert?

6. Bij de keuring van een zeer grote partij speelgoedauto's heeft men de volgende keuringsvoorschrift:

- neem aselect 10 stuks uit de partij;
- zijn hiervan 3 of meer ondeugdelijk, (niet voldoen aan de gestelde eisen) keur de partij af;
- zijn er 1 of 2 exemplaren ondeugdelijk, neem nog een aselecte steekproef van 10 stuks; keur de partij goed als er in beide steekproeven tezamen ten hoogste 3 exemplaren ondeugdelijk zijn;
- zijn er in de eerste steekproef 0 exemplaren ondeugdelijk, keur de partij goed.

Wat is de kans dat een partij met 20% ondeugdelijke exemplaren wordt afgekeurd?

7. De kans dat een fluorescerende lamp een levensduur heeft van minstens 500 uur bedraagt 85%. Bereken de kans dat bij 20 van deze lampen men:
a. precies 18 lampen vindt met een levensduur van ten minste 500 uur;
b. tenminste 15 lampen vindt met een levensduur van ten minste 500 uur;
c. ten hoogste 2 lampen vindt die geen levensduur van ten minste 500 uur hebben.

8. Gemiddeld genomen treedt er in een fabriek één keer per 50 werkdagen een storing op in een bepaalde machine. Hoe groot is de kans dat er in een aaneengesloten periode van 10 werkdagen twee keer een storing in deze machine optreedt?

9. Op een klein vliegveld arriveren per uur gemiddeld 3 vliegtuigen. Hoe groot is de kans dat er:
a. in een periode van 2 uur hoogstens 2 vliegtuigen landen?
b. in een periode van een $\frac{1}{2}$ uur minstens 2 vliegtuigen landen?
c. in een periode van 3 uur minstens 5 maar hoogstens 8 vliegtuigen landen?

10. Een verhuurbedrijf heeft 10 auto's, die per dag worden verhuurd. Gemiddeld over een lange periode blijken er 7,5 aanvragen voor verhuur per dag te zijn.
Hoe groot is de kans dat op een dag alle wagens zijn verhuurd?

11. Een transportbedrijf heeft 2 grote hijskranen die per werkdag gehuurd kunnen worden. Het aantal aanvragen per dag bedraagt gemiddeld 1,5. Onder een werkdag wordt verstaan van 08.00 uur tot en met 17.00 uur.
a. Hoe groot is de kans dat er om 11.00 uur nog geen aanvraag is binnengekomen?
b. Welk percentage van de dagen zijn beide hijskranen thuis?
c. Welk percentage van de dagen zijn beide hijskranen verhuurd?

12. Een drukker wenst tegen zijn klanten de volgende bewering te gebruiken:
'De kans dat er meer dan m drukfouten op een willekeurige pagina voorkomen, is kleiner dan 1%'. Als het aantal drukfouten per pagina een Poisson-verdeling volgt, met een gemiddelde van 3, welke waarde van m zal de drukker dan moeten kiezen in zijn bewering?

13. Een keuringsambtenaar onderzoekt een aselecte steekproef van drie broodroosters uit een partij van 24. Als de partij zes broodroosters bevat met kleine gebreken, wat is dan de kans dat de keuringsambtenaar zal vinden:
a. geen defecte broodroosters?
b. slechts één defect broodrooster?
c. ten minste twee defecte broodroosters?

14. Een bingo-spelletje wordt gespeeld met 35 getallen (van 1 t/m 35).
a. Wat is de kans dat een deelnemer van 10 getrokken getallen er 0 goed heeft?
b. Wat is de kans dat de deelnemer 5 van de 10 getallen goed heeft?

6 Continue kansverdelingen

6.1 Inleiding

In dit hoofdstuk bespreken we de continue kansverdelingen, met als belangrijkste continue (kans)verdeling de normale verdeling.
Een continu verdeelde kansvariabele kan elk reëel getal op een bepaald interval aannemen. Dit in tegenstelling tot een discreet verdeelde kansvariabele die alleen gehele getalswaarden kan aannemen of een geselecteerd aantal waarden op een zeker interval.
Een *continue kansverdeling* ontstaat als we gaan meten. Hierbij kan in principe elke reële getalswaarde in een bepaald interval voorkomen. Het hangt van de nauwkeurigheid van het meetinstrument af, hoe nauwkeurig (in hoeveel decimalen) de meetuitkomst wordt weergeven. In zijn algemeenheid kunnen we zeggen dat, als een meetwaarde in decimalen kan worden weergegeven, we te maken hebben met een continue verdeling. Soms wordt een continue variabele echter wel in gehele getallen weergeven, maar als het meetinstrument wat nauwkeuriger was, waren ook decimale waarden mogelijk. Het gewicht van een bepaald product kan men weergeven in gehele grammen (al dan niet na afronding), maar ook in cijfers achter de komma. Dus gewicht is een continue variabele. Tentamencijfers zijn meestal gehele getallen. In die zin hebben we te maken met een discrete variabele. Maar we kunnen tentamencijfers ook zien als afgeronde reële getallen. Daardoor kunnen we de verdeling van tentamencijfers toch beschouwen als een continue kansverdeling.
Net als in hoofdstuk 5 bij discrete kansvariabelen willen we een theoretisch model van een aantal belangrijke continue verdelingen maken. Feitelijk is de benaming kansfunctie, die in hoofdstuk 5 voor discrete kansvariabelen is geïntroduceerd, in dit geval niet hanteerbaar. Om het wiskundig model van een continue kansverdeling te kunnen beschrijven, gebruiken we de benaming *kansdichtheid* (of *kansdichtheidsfunctie*).

Het begrip kansdichtheid

Dat een kansdichtheid een positieve uitkomst moet hebben ligt voor de hand (een kans is per definitie positief). We weten inmiddels dat een continue kansvariabele op een zeker

interval (bijvoorbeeld op $[a, b]$) oneindig veel waarden kan hebben. Dat de som van alle oneindig-veel kansen gelijk aan 1 moet zijn, kunnen we praktisch gezien niet meer als som maar wel als integraal schrijven. Vandaar:

Definitie

Voor een kansdichtheid $f(x)$ van een continue kansvariabele X, die waarden x aan kan nemen op het interval $[a, b]$, geldt in het algemeen:

1. $f(x) \geq 0$ voor alle waarden x van X

2. $\int_a^b f(x)dx = 1$

Opmerking

a. We merken op dat *a* en/of *b* oneindig klein respectievelijk groot kunnen zijn.

b. We herinneren eraan dat voor zover het de naam van een kansvariabele betreft deze met een hoofdletter geschreven wordt (X) en wanneer we de waarde van de variabele bedoelen, schrijven we een kleine letter (x).

c. De kans dat X exact de waarde x (willekeurig) op het interval $[a, b]$ aanneemt, is zeer klein (vrijwel 0), omdat op het interval oneindig veel mogelijke waarden liggen.

De definitie houdt in dat het oppervlak onder de grafiek van een kansdichtheid gelijk aan 1 moet zijn. Deeloppervlakken kunnen we identificeren met kansen:

$$P(c < X < d) = \int_c^d f(x)dx \tag{6.1}$$

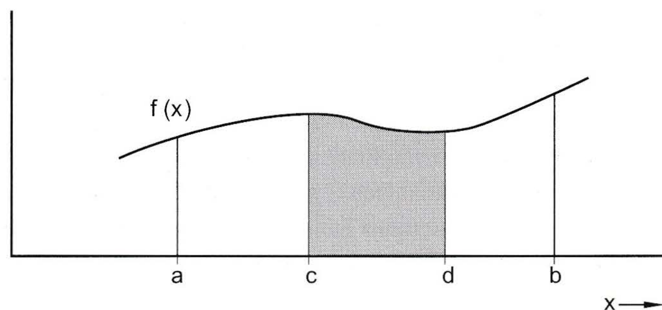


Fig. 6.1 $P(c < X < d) = \int_c^d f(x)dx$

Opdracht

Bedenk waarom voor een continue kansvariabele X geldt: $P(c \leq X \leq d) = P(c < X < d)$

6.2 Verwachtingswaarde en variantie van een continue kansverdeling

Net als bij een discrete kansvariabele (zie hoofdstuk 5) bestaan er formules voor de verwachtingswaarde (gemiddelde) en variantie (of standaardafwijking) van een continue kansvariabele. Deze formules vertonen een zekere analogie.
Voor de verwachtingswaarde van een continue kansvariabele geldt:

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx \tag{6.2}$$

We kunnen deze formule begrijpen als we de formule voor de verwachtingswaarde van een discrete kansvariabele tevoorschijn halen: $E(K) = \sum_{i=1}^{n} k_i f(k_i)$ en bedenken dat de variabele X, in tegenstelling tot K, een interval met oneindig veel reële getallen doorloopt.
Als X slechts waarden tussen a en b doorloopt gaat formule (6.2) over in:
$E(X) = \int_a^b x \cdot f(x)dx$
Voor de *variantie* van een continue verdeling geldt, net als voor een discrete verdeling:

$$var(X) = E(X^2) - \{E(X)\}^2$$

Voor $E(X^2)$ geldt, in analogie met de formules (5.3a) en (5.4a) uit hoofdstuk 5 $\left(\text{waarin } E(K^2) = \sum_{i=1}^{n} f(k_i) \cdot k_i^2 \text{ voorkomt}\right)$, ook weer na overgang op een integraal:

$$E(X^2) = \int_{-\infty}^{\infty} x^2 \cdot f(x)dx$$

Gebruikmakend van formule (6.2) geldt voor de variantie van een continue kansvariabele dus:

$$var(X) = \int_{-\infty}^{\infty} x^2 \cdot f(x)dx - \left(\int_{-\infty}^{\infty} x \cdot f(x)dx\right)^2 \tag{6.3}$$

$$= \int_{-\infty}^{\infty} x^2 \cdot f(x)dx - \mu^2 \tag{6.3a}$$

De formules (6.1) t/m (6.3a) zullen we in de rest van dit hoofdstuk gebruiken.
In dit hoofdstuk zullen we nu drie continue kansverdelingen bespreken, namelijk:

- de uniforme of rechthoekige continue verdeling;
- de normale verdeling;
- de negatief-exponentiële verdeling.

De normale verdeling is verreweg de belangrijkste van alle theoretische kansverdelingen. Deze verdeling vormt de grondslag van de klassieke statistische toetsings- en schattingstheorie (zie hoofdstuk 7, 8 en 9). Allereerst zullen we de uniforme of rechthoekige continue verdeling behandelen.

6.3 Uniforme- of rechthoekige continue verdeling

Zoals we in de inleiding gezien hebben, geldt voor een continue kansvariabele dat elke uitkomst, in een bepaald interval mogelijk is. Voor de uniforme of rechthoekige continue verdeling geldt daarbij dat het optreden van elke uitkomst dezelfde kans bezit. De kansdichtheid $f(x)$ is voor alle waarden in de uitkomstenruimte(interval) constant. Als voorbeeld nemen we een 'rad van avontuur'.

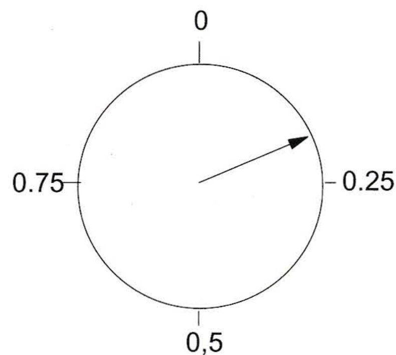


Fig. 6.2 'Rad van avontuur'

De pijl in figuur 6.2 draait op een as in het middelpunt van een cirkelvormige schijf. De pijl beweegt in de richting van de wijzers van de klok met een willekeurige beginsnelheid. Na enige tijd komt de pijl tot stilstand. We bekijken de plaats op de rand waar de pijl blijft stilstaan. Omdat de beginsneldheid willekeurig is, is elk punt op de rand waar de pijl tot stilstand komt even waarschijnlijk. De plaats waar de pijl stopt is te identificeren met een willekeurig reëel getal tussen 0 en 1. Elk getal tussen 0 en 1 is even waarschijnlijk. De som van alle mogelijke kansen, dus de totale kans 1, wordt uniform verdeeld over alle mogelijke waarden van X in het interval $[0,1]$. Een dergelijke kansvariabele X noemen we

een *standaard uniforme kansvariabele*. We zeggen ook wel dat de kansvariabele X een rechthoekige of uniforme verdeling volgt op [0,1].
De formule voor de kansdichtheid van X luidt in dit geval:

$$\begin{cases} f(x) = 1 & \text{voor } 0 \leq x \leq 1 \\ f(x) = 0 & \text{elders} \end{cases}$$

Hiermee zeggen we dat elke waarde van X even waarschijnlijk is.

Opdracht

Controleer voor bovenstaande functie dat aan de definitie van een kansdichtheid is voldaan.

We kunnen op basis van het gegeven voorbeeld wat algemener definiëren:

Definitie
Een continue kansvariabele X die alle waarden x op het interval $[a, b]$ kan aannemen met even grote waarschijnlijkheid, dat wil zeggen met *constante* kansdichtheid, heet een *uniforme continue kansvariabele.* De formule voor de kansdichtheid hiervoor luidt:

$$\begin{cases} f(x) = \dfrac{1}{b-a} & \text{voor } a \leq x \leq b \\ f(x) = 0 & \text{elders} \end{cases} \tag{6.4}$$

In figuur 6.3 is de grafiek van de kansdichtheid uit formule (6.4) weergegeven.

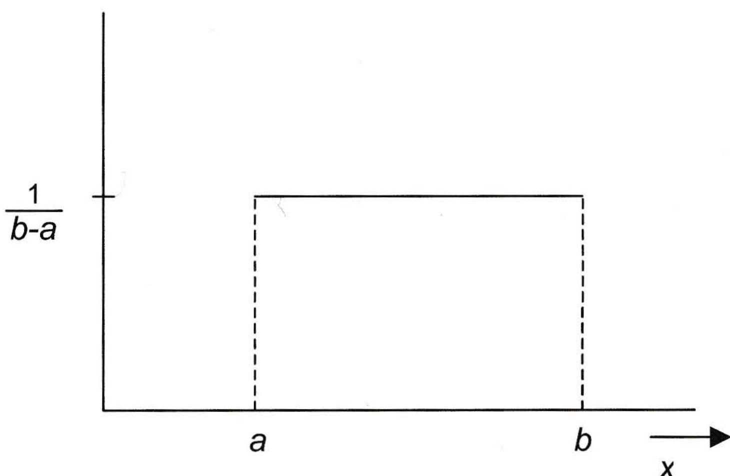


Fig. 6.3 Kansdichtheid van een uniform verdeelde X

Aan de symmetrie in de grafiek van de kansdichtheid zien we dat voor de verwachtingswaarde van een uniform verdeelde kansvariabele geldt: $E(X) = \frac{1}{2}(b + a)$. Het gemiddelde ligt precies halverwege tussen a en b. We kunnen dit bewijzen door formule (6.2) toe te passen.

Voorbeeld 1

Als we het waardeninterval $[a, b]$ beschouwen voor de continue rechthoekig-verdeelde kansvariabele X, met $f(x) = \dfrac{1}{b-a}$, krijgen we voor de verwachtingswaarde van X:

$$\begin{aligned} E(X) &= \int_a^b x \cdot \frac{1}{b-a} dx = \frac{1}{b-a}\left[\frac{1}{2}x^2\right]_{x=a}^{x=b} \\ &= \frac{1}{2(b-a)}(b^2 - a^2) \\ &= \frac{(b-a)(b+a)}{2(b-a)} = \frac{1}{2}(b+a) \end{aligned}$$

Het gemiddelde of de verwachtingswaarde van een op het interval $[a, b]$ rechthoekig continue verdeelde variabele X is dus:

$$E(X) = \mu = \frac{1}{2}(b + a)$$

Voor de berekening van de variantie kunnen we niet zonder formule (6.3):

Voorbeeld 2

Voor de continue op het interval $[a, b]$ rechthoekige-verdeelde verdeelde variabele X berekenen we de variantie als volgt:

$$\begin{aligned} E(X^2) &= \int_a^b x^2 \cdot \frac{1}{b-a} dx = \frac{1}{3(b-a)} \cdot (b^3 - a^3) \\ &= \frac{(b-a)(b^2 + ab + b^2)}{3(b-a)} \\ &= \frac{b^2 + ab + b^2}{3} \end{aligned}$$

Er geldt dus:

$$\begin{aligned} var(X) &= \frac{b^2+ab+b^2}{3} - \left(\frac{1}{2}(b+a)\right)^2 \\ &= \frac{1}{12}(b-a)^2 \end{aligned}$$

In de laatste regel is een wiskundige herleiding toegepast.

De zojuist ontwikkelde formules kunnen algemeen gebruikt worden bij uniforme of rechthoekig-continu-verdeelde kansvariabelen.

Voorbeeld 3

De verwachting en de variantie van een continue rechthoekig verdeelde variabele op het interval [3,13] is :

$E(X) = \mu = \frac{1}{2}(a+b) = \frac{1}{2}(3+13) = 8$

$var(X) = \frac{1}{12}(b-a)^2 = \frac{1}{12}(13-3)^2 = \frac{100}{12}$

6.4 De normale verdeling

Vele verschijnselen uit de natuur, zoals bijvoorbeeld de frequentieverdeling van lengtes van geproduceerde assen of het gewicht van mensen, geven bij grafische weergave een histogram waarvan de vorm (weergegeven door de zogenaamde ideale kromme) ongeveer klokvormig is. Deze grafieken worden vaak benaderd door een continue kromme, die ééntoppig en symmetrisch is. Deze kromme is voor het eerst ontdekt (onafhankelijk van elkaar) door de Franse wiskundigen De Moivre en Laplace. De Duitser Carl Friedrich Gauss (1777 - 1885) gaf het belang van de kromme weer, door deze in verband te brengen met de foutentheorie van fysische metingen.
Doordat de kromme in de praktijk veelvuldig voorkomt, wordt zij de 'normale verdeling' genoemd. Men moet aan het woord 'normaal' echter geen specifieke betekenis toekennen, in de zin dat een verdeling die niet normaal is, beschouwd zou moeten worden als abnormaal ofwel 'malafide'.
Daarnaast wordt de kromme ook vaak naar zijn ontdekkers genoemd, namelijk de Gauss-kromme, of verdeling van Gauss, of verdeling van De Moivre.
Naast de al genoemde praktische zin en het belang van de normale verdeling (meetfouten en verschijnselen in de natuur), kunnen andere kansverdelingen (zoals de binomiale en de Poisson-verdeling) onder bepaalde omstandigheden benaderd worden door een normale verdeling. We komen hierop terug aan het eind van dit hoofdstuk.
In figuur 6.4 is de frequentieverdeling weergegeven van de lengtemetingen van 1000 aselect gekozen mannen. De klassenbreedte is 2 cm. Vertikaal is het aantal mannen per klasse weergegeven.

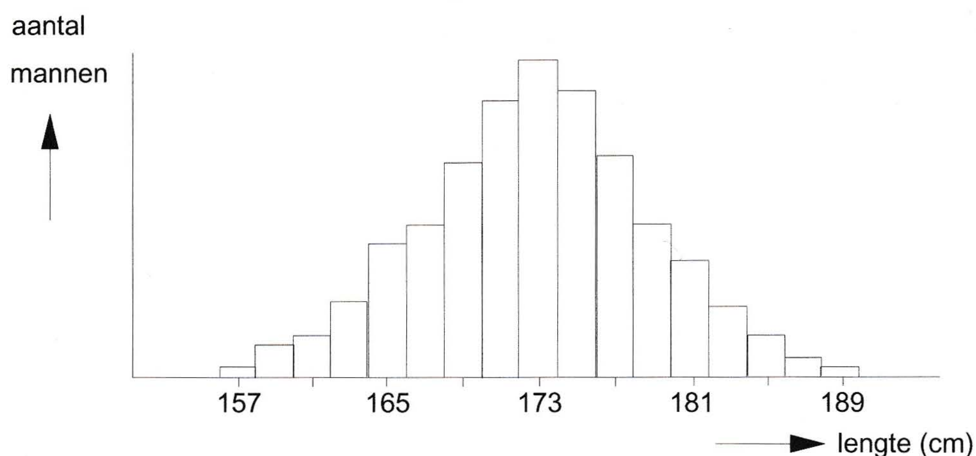


Fig. 6.4 De lengteverdeling van 1000 mannen (klassenbreedte = 2)

Uit de afbeelding blijkt dat de klasse 172 - 174 de modale klasse is. Dit is de klasse met 173 als klassenmidden. Op grond van de vorm van het histogram kunnen we al vermoeden dat de lengten normaal verdeeld zijn.

In plaats van de absolute aantallen, kunnen ook de relatieve aantallen vertikaal worden uitgezet. De relatieve frequentie is, zoals we in hoofdstuk 4 gezien hebben, gelijk aan de kans dat een waarneming in een bepaalde klasse valt (en komt, zoals we in hoofdstuk 5 gezien hebben overeen met de *kansfunctie*). In figuur 6.5 is de kansfunctie uitgezet van voorbeeld 4, maar nu bij een steekproef van 3000 mannen. De klassenbreedte is daarbij verkleind tot 1 cm.

De kans dat een man een lengte heeft van 173 cm (dus in de klasse 172,5 - 173,5 met klassenmidden 173 cm valt), is het grootst.

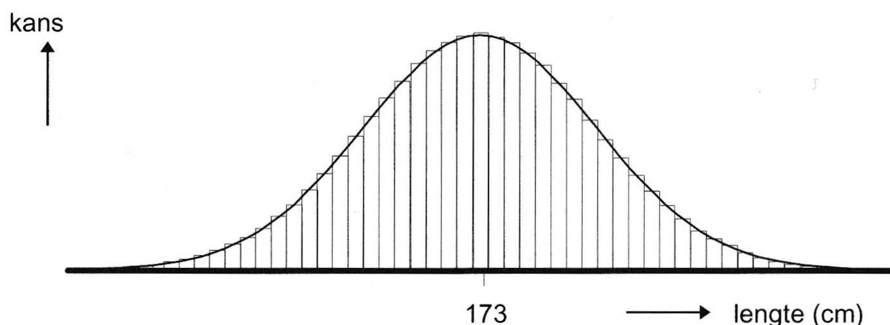


Fig. 6.5 De lengteverdeling van 3000 mannen (klassenbreedte = 1)

De 'trapjeskromme' kunnen we benaderen door een vloeiende kromme (zie figuur 6.5).

De vloeiende kromme die hierdoor ontstaat, vertoont ongeveer een (kerk)klokvorm. Door de klassenbreedte steeds meer te verkleinen (naderend tot 0) en het aantal metingen te vergroten, kan de vloeiende kromme steeds beter worden geconstrueerd en is de klokvorm des te beter te benaderen. De discrete 'trapjes'-verdeling gaat hierbij over in een continue verdeling.
De op deze wijze ontstane kromme, kan blijkens de ontdekking van Gauss, goed benaderd worden door de wiskundige functie:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{6.5}$$

Dit is de formule van de *kansdichtheid van de normale verdeling*behorend bij de kansvariabele X.
Bewezen kan worden dat formule (6.5) voldoet aan de definitie van een kansdichtheid: het oppervlak onder de kromme is, ongeacht de waarde van μ en σ gelijk aan 1.
In figuur 6.6 is de grafiek van deze kansdichtheid weergegeven.

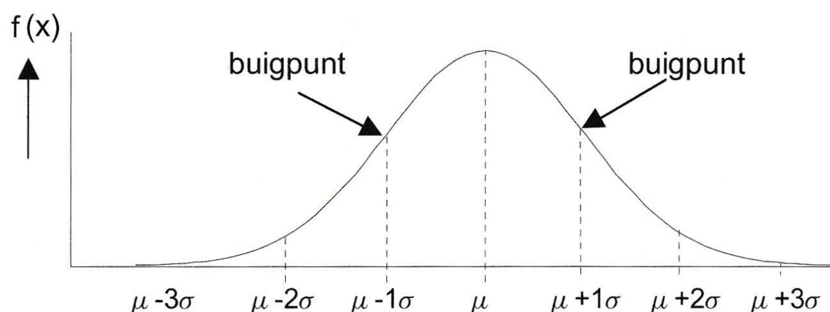


Fig. 6.6 Kansdichtheid van de normale verdeling

Zowel uit de formule als uit de figuur blijkt dat de normale verdeling volledig wordt bepaald door de *parameters* μ en σ. De normale verdeling blijkt klokvorming te zijn en symmetrisch om $x = \mu$. Verder kan bewezen worden dat er buigpunten zijn bij de punten met x-coördinaat $x = \mu + \sigma$ en $x = \mu - \sigma$.
Passen we formule (6.2) en formule (6.3) toe met de kansdichtheid uit formule (6.5) dan blijkt dat de verwachtingswaarde $E(X)$ gelijk aan μ is en de variantie $var(X)$ gelijk aan σ^2 is, dus de standaardafwijking is gelijk aan σ (voor het bewijs verwijzen we naar de wiskundeboeken). Kortom: de parameters van de normale verdeling zijn de verwachtingswaarde μ (= het gemiddelde) en de standaardafwijking σ.

Opdracht

Ga na dat de grafiek van de normale verdeling breed en laag is bij een grote waarde van σ (veel spreiding) en smal en steil bij een kleine waarde van σ (weinig spreiding).

In principe loopt het waardenbereik van X van $-\infty$ tot $+\infty$. In de praktijk komt het echter niet voor dat de uitkomsten van metingen een dergelijk interval doorlopen. Bijvoorbeeld mannen kleiner dan 0 meter of langer dan 3 meter is natuurlijk onzin. Het blijkt dat $f(x)$ voor zulke onwaarschijnlijke uitkomsten praktisch nul is en de theoretische benadering van het werkelijk geval door $f(x)$ geen belemmering vormt. Voor elke normale verdeling geldt dat van de totale oppervlakte:

- 68,2% ligt tussen de grenzen $\mu - 1\sigma$ en $\mu + 1\sigma$;
- 95,4% ligt tussen de grenzen $\mu - 2\sigma$ en $\mu + 2\sigma$;
- 99,7% ligt tussen de grenzen $\mu - 3\sigma$ en $\mu + 3\sigma$.

Deze feiten zullen we spoedig aantonen. Ze zijn weergegeven in figuur 6.7.

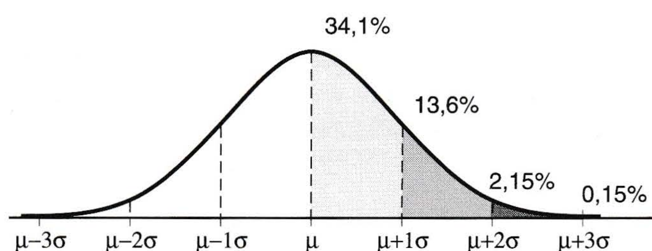


Fig. 6.7 De normale verdeling

We zien hierin dat binnen de 2σ-grenzen, de gebruikelijke aanduiding voor het interval $[\mu - 2\sigma, \mu + 2\sigma]$, ruim 95% van alle waarnemingsuitkomsten liggen.
Binnen de 3σ-grenzen (interval $[\mu - 3\sigma, \mu + 3\sigma]$ ligt 99,7%. Hebben we bijvoorbeeld 300 uitkomsten, dan zal er slechts één buiten de 3σ-grenzen vallen.

Voorbeeld 4

Stel dat de lichaamslengten van mannen normaal verdeeld zijn met $\mu = 174$ cm en $\sigma = 7$ cm, dan ligt tussen de 2σ-grenzen (174 ± 14 cm) 95,4% van de waarnemingen. Men kan dan ook zeggen dat 95,4% van alle mannen een lengte heeft tussen 160 en 188 cm. Weer anders gezegd: de kans om een lengtewaarde tussen 160 en 188 cm te vinden is 0,954 ofwel 95,4%. Dit laatste kunnen we beknopt weergeven als:
$P(160 < X < 188) = 0{,}954$.

Men kan ook twee willekeurige grenswaarden nemen bijvoorbeeld $X = a$ en $X = b$. Zoals in de inleiding van dit hoofdstuk reeds is gezegd, geldt voor elke continue kansverdeling dat $P(a < X < b)$ gelijk is aan dat deel van de totale oppervlakte onder de kromme dat tussen a en b ligt (zie figuur 6.8). De totale oppervlakte onder de kromme is 1 of 100%.

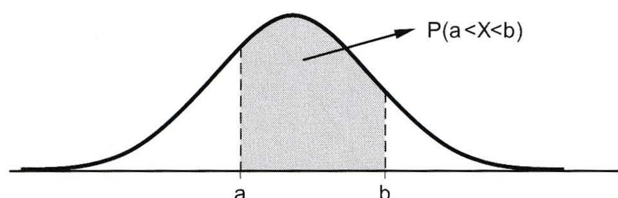


Fig. 6.8 Oppervlakte = kans

Opdracht

Toon aan dat $P(a < X < b) = P(X > a) - P(X > b)$

Om $P(a < X < b)$ te berekenen, moeten we $f(x)$ integreren van a naar b. Deze integraal is niet eenvoudig te bepalen. Tabelleren is ook moeilijk want er bestaan in principe oneindig veel waarden voor μ en σ. Er bestaan dus ook oneindig veel normale verdelingen. Door een eenvoudige transformatie kan elke willekeurige normale verdeling herleid worden tot de zogenaamde *standaardnormale verdeling*.

6.4.1 Standaardnormale verdeling (u-verdeling)

De transformatieformule voor de transformatie van een willekeurige normale verdeling (met parameters μ en σ) naar de standaardnormale verdeling is:

$$U = \frac{X - \mu}{\sigma} \text{ dus ook } u = \frac{x - \mu}{\sigma} \tag{6.6}$$

Alle x-waarden worden getransformeerd naar u-waarden, door x te verminderen met het gemiddelde μ en daarna te delen door de standaardafwijking σ. In afbeelding 6.7 is het verband tussen x en u schematisch weergegeven.
Men kan bewijzen dat de u-waarden weer normaal verdeeld zijn. Uit de figuur leiden we direct af dat voor de gestandaardiseerde eenheden geldt dat $\mu_U = 0$ en $\sigma_U = 1$. Door de transformatie volgens formule (6.6) wordt een kans als $P(X > a)$ getransformeerd naar $P\left(U > \frac{a - \mu}{\sigma}\right)$. In figuur 6.8 is de kans $P(X > a)$ gelijk aan het oppervlak onder de grafiek van de normale verdeling, *rechts* gelegen van de lijn $x = a$. Daarom noemt men dit een *rechteroverschrijdingskans*. In de standaardnormale verdeling is $P\left(U > \frac{a - \mu}{\sigma}\right)$ uiteraard ook weer een rechteroverschrijdingskans.

Opmerking

In het algemeen wordt de normale verdeling met gemiddelde μ en variantie σ^2, weergegeven door $N(\mu, \sigma^2)$. Als X normaal verdeeld is, schrijven we $X \sim N(\mu, \sigma^2)$. De standaardnormale verdeling wordt geschreven als $N(0, 1^2)$.

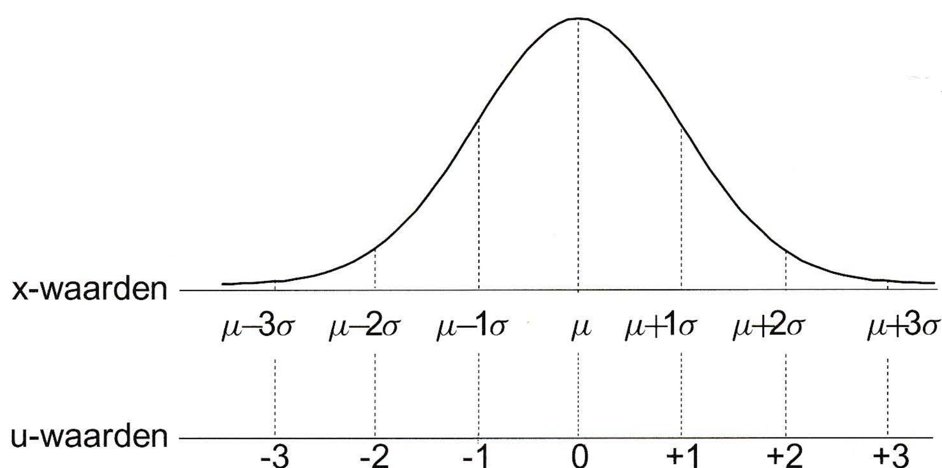


Fig. 6.9 De oorspronkelijke schaal (x) en de gestandaardiseerde schaal (u)

Een groot aantal overschrijdingskansen bij een standaardnormale verdeling is in een tabel weergegeven. Doordat de standaardnormale verdeling volledig symmetrisch is om $u = 0$, wordt alleen de rechteroverschrijdingskans in de tabel gegeven. In tabel B1 zijn de rechteroverschrijdingskansen gegeven. In figuur 6.10 is een voorbeeld gegeven van een rechteroverschrijdingskans.

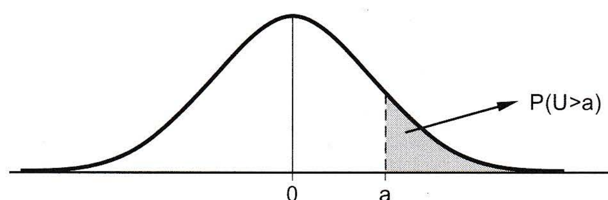


Fig. 6.10 Rechteroverschrijdingskans ($P(U > a)$)

Vanwege de symmetrie van de normale verdeling geldt dat: $P(U < -a) = P(U > +a)$.
Dit betekent dat de linkeroverschrijdingskans $P(U < -a)$ gelijk is aan de rechteroverschrijdingskans $P(U > +a)$.
De kans op een u-waarde kleiner dan een bepaalde negatieve a waarde (linkeroverschrijdingskans) kunnen we daarom vinden door de rechteroverschrijdingskans van die positieve a-waarde op te zoeken in de tabel.

6.4.2 De tabel voor de standaardnormale verdeling

We zullen nu uitleggen hoe de tabel van de standaardnormale verdeling te gebruiken is.

Voorbeeld 5

De lichaamslengte van volwassen mannen is normaal verdeeld met parameters $\mu = 174$ en $\sigma = 7$ (men schrijft dus $X \sim N(174, 7^2)$). Wat is de kans dat een willekeurige man langer is dan 176,4 cm?

Oplossing

We moeten bepalen $P(X > 176{,}4)$. We transformeren nu de normaal verdeelde variabele X naar de standaardnormaal verdeelde variabele U. Tegelijk transformeren we de waarde $x = 176{,}4$ naar de bijbehorende u-waarde van de standaardnormale verdeling door de transformatie:

$$u = \frac{x - \mu}{\sigma} = \frac{176{,}4 - 174}{7} \approx 0{,}34$$

Vervolgens zoeken we de rechteroverschrijdingskans bij $u = 0{,}34$ op in tabel B1.
Het aflezen gaat als volgt: In de voorkolom zoeken we het eenhedencijfer en het eerste decimale cijfer, dus 0,3, en gaan vandaar naar rechts tot de kolom waarboven het tweede decimale cijfer – dus 4 – staat.

u	.01	.02	.03	.04	...
0.0					
0.1					
0.2					
0.3				3669	
0.4					
enz					

We vinden het getal 3669, hetgeen wil zeggen: $P(U > 0{,}34) = 0{,}3669$ ofwel 36,69%. Dit is de gevraagde kans op een lengte van meer dan 176,4 cm.

Opmerking

Het verdient aanbeveling bij de berekening van de kansen voor een normale verdeling steeds een schetsje te tekenen en de gevraagde oppervlakte te arceren. We zien dan meteen hoe de gevraagde kans moet worden verkregen, door:

- rechtstreeks aflezen (bijvoorbeeld bij $P(U > a)$, met $a > 0$);
- aftrekken (bijvoorbeeld bij $P(a < U < b) = P(U > b) - P(U > a)$, met a en b > 0);
- optellen (bijvoorbeeld bij $P(a < U < b)$ met negatieve a en positieve b; in dat geval kunnen we schrijven $P(a < U < b) = P(a < U < 0) + P(0 < U < b)$).

6.4.3 Rekenvoorbeelden

We zullen het gebruik van tabel B1 illustreren met een aantal voorbeelden.

Voorbeeld 6

Van een bepaald product wordt per week gemiddeld $\mu = 50$ ton omgezet met een standaardafwijking van $\sigma = 5$ ton. Aangenomen mag worden dat de omzetten normaal verdeeld zijn. Bereken achtereenvolgens:

a. de kans op een omzet groter dan 57 ton;
b. de kans op een omzet kleiner dan 53 ton;
c. de kans op een omzet tussen 44 en 48 ton;
d. boven welke grens zal 15% van de omzetten liggen?

Oplossing

a. $P(X > 57) = P\left(U > \dfrac{57-50}{5}\right) = P(U > 1{,}40)$

Opzoeken in tabel B1 leidt tot $P(U > 1{,}40) = 0{,}0808$
De kans dat de omzet groter is dan 57 ton bedraagt dus 8,08%

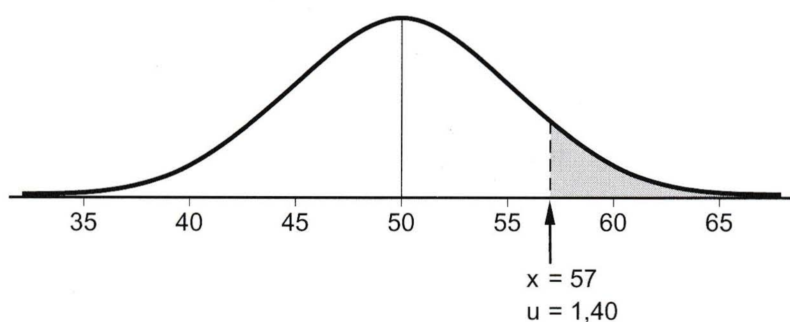


Fig. 6.11a

b. $P(X < 53) = 1 - P(X > 53)$

$P(X > 53) = P\left(U > \dfrac{53-50}{5}\right) = P(U > 0{,}60)$

Volgens tabel B1 is $P(U > 0{,}60) = 0{,}2743$
De gevraagde kans is dus $1 - 0{,}2743 = 0{,}7257$

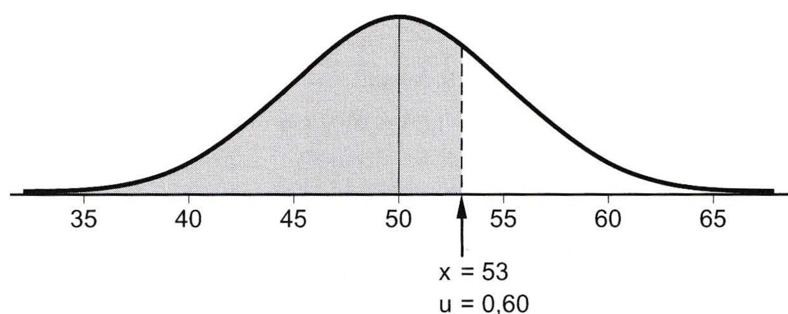


Fig. 6.11b

c. We berekenen hier zowel de linkeroverschrijdingskans voor 44 ton, als voor 48 ton:

$P(X < 44) = P\left(U < \dfrac{44-50}{5}\right) = P(U < -1{,}20)$

$P(U < -1{,}20) = P(U > 1{,}20) = 0{,}1151$

$P(X < 48) = P\left(U < \dfrac{48-50}{5}\right) = P(U < -0{,}40)$

$P(U < -0{,}40) = P(U > 0{,}40) = 0{,}3446$

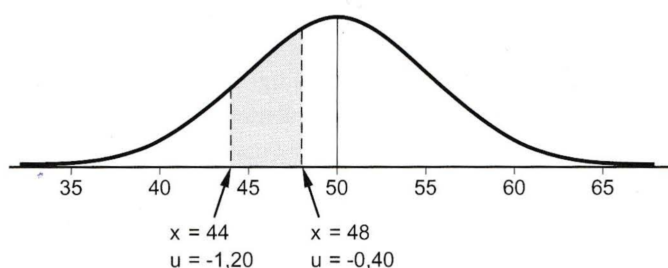


Fig. 6.11c

c. (vervolg)

De gevraagde kans is:

$P(44 < X < 48) = P(X > 44) - P(X > 48) = 0{,}3446 - 0{,}1151 = 0{,}2295$

d. Hier hebben we te maken met de omgekeerde situatie, namelijk de overschrijdingskans is gegeven terwijl de waarde voor x niet bekend is maar juist gevraagd wordt. We zoeken in tabel B1 welke u-waarde een rechteroverschrijdingskans heeft van 0,1500. Deze u-waarde is niet exact te vinden. We nemen die u-waarde uit de tabel die een overschrijdingskans heeft die het dichtste bij 0,1500 ligt.

We vinden twee kandidaten:

$u = 1{,}04 \Rightarrow P = 0{,}1492$

$u = 1{,}03 \Rightarrow P = 0{,}1515$

De overschrijdingskans bij $u = 1{,}04$ ligt dichter bij $P = 0{,}150$ dan de kans bij $u = 1{,}03$.

In de transformatieformule $u = \dfrac{x-\mu}{\sigma}$ vullen we daarom voor $u = 1{,}04$ in.
Met $\mu = 50$ en $\sigma = 5$ is x nu op te lossen:
$1{,}04 = \dfrac{x-50}{5} \Rightarrow x = 50 + 5 \times 1{,}04 = 55{,}2$ ton

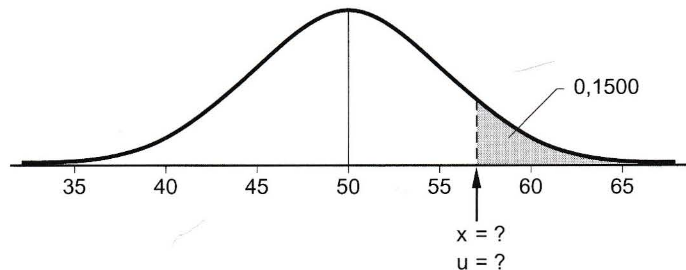


Fig. 6.11d

6.5 Benadering van een discrete verdeling door een normale verdeling

We komen nu terug op een eerder gemaakte opmerking. Soms is het mogelijk het rekenwerk voor een discreet verdeelde variabele te vereenvoudigen door de normale verdeling te gebruiken. Maar dan moet wel aan een aantal voorwaarden voldaan zijn.

6.5.1 Benadering van een binomiale verdeling door een normale verdeling

De binomiale verdeling met de parameters n en p tendeert voor grotere waarden van n naar een symmetrische verdeling. We kunnen dit verschijnsel zelf onderzoeken door naar de tabel B2 te kijken. Hoe groter n is, hoe meer symmetrisch de binomiale verdeling is. De symmetrie is des te sterker als p dichter bij $\frac{1}{2}$ ligt. Dit verschijnsel leidt tot de vraag of het niet mogelijk is een binomiale verdeling onder bepaalde voorwaarden te benaderen door een normale verdeling.

Het ligt voor de hand om als *eerste* voorwaarde te stellen dat de normale verdeling, waarmee we de binomiale verdeling met de parameters n en p willen benaderen, hetzelfde gemiddelde en dezelfde standaardafwijking heeft als de binomiale verdeling. Dit betekent dat die normale verdeling een gemiddelde $\mu = np$ en een standaardafwijking $\sigma = \sqrt{np(1-p)}$ heeft (zie formule (5.6) en formule (5.8) uit hoofdstuk 5) heeft.

Wat de *tweede* voorwaarde betreft: de normale verdeling is een continue kansverdeling, maar de binomiale verdeling is een discrete kansverdeling. Willen we een (discrete) binomiale verdeling benaderen door een (continue) normale verdeling, dan zal als voorwaarde gesteld moeten worden dat de zogenaamde *continuïteitscorrectie* wordt toegepast. Wat we hieronder verstaan blijkt uit figuur 6.12.

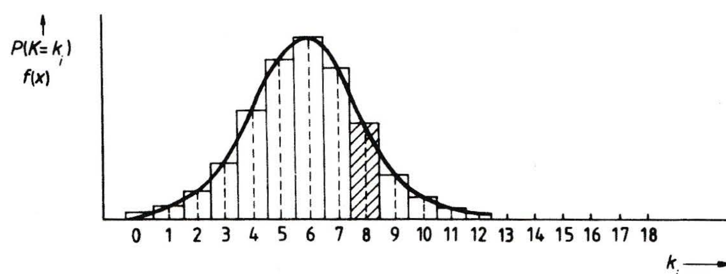


Fig. 6.12 Benadering door een normale verdeling

In figuur 6.12 bestaat de discrete binomiale verdeling met de parameters $n = 18$ en $p = \frac{1}{3}$ en de continue normale verdeling met de parameters $\mu = 18 \cdot \frac{1}{3} = 6$ en $\sigma = \sqrt{18 \cdot \frac{1}{3} \cdot \frac{2}{3}} = 2$. In deze figuur is de 'binomiale' kans $P(K = 8)$ gelijk aan de oppervlakte van de kolom behorend bij $K = 8$ in de binomiale verdeling en is de 'normale' kans $P(7{,}5 < X < 8{,}5)$ gelijk aan de oppervlakte van de figuur begrensd door de Gauss-kromme, de lijnen $x = 7{,}5$ en $x = 8{,}5$ en de x-as.

Opmerking

Om duidelijk te accentueren dat we overgaan van een discrete naar een continue verdeling, veranderen we de naam van de variabele van K in X.

Uit de figuur blijkt dat de eerstgenoemde oppervlakte goed benaderd kan worden door de laatstgenoemde oppervlakte. Met andere woorden: de 'binomiale' kans $P_B(K = 8)$ met $n = 18$ en $p = \frac{1}{3}$ kan benaderd worden door de 'normale' kans $P_N(7{,}5 < X < 8{,}5)$ met $\mu = 6$ en $\sigma = 2$.

Opdracht

Controleer deze uitspraak door de 'binomiale' kans $P_B(K = 8)$ en de 'normale' kans $P_N(7{,}5 < X < 8{,}5)$ met dezelfde gegevens voor n en p te berekenen.

Willen we nu bijvoorbeeld de 'binomiale' kans $P_B(5 \leq K \leq 7)$ berekenen, dan kan dit door te bedenken dat:

$$\begin{aligned} P_B(5 \leq K \leq 7) &= P_B(K=5) + P_B(K=6) + P_B(K=7) \\ &= P_N(4{,}5 < X < 5{,}5) + P_N(5{,}5 < X < 6{,}5) + P_N(6{,}5 < X < 7{,}5) \\ &= P_N(4{,}5 < X < 7{,}5) \end{aligned}$$

Opdracht

Bereken voor het geval van figuur 6.12 de 'binomiale' kans $P_B(5 \leq K \leq 7)$ en de normale benadering $P_N(4{,}5 < X < 7{,}5)$. Ga vervolgens door berekening van $P_N(5 < X < 7)$ na dat weglating van de continuïteitscorrectie een minder goede benadering oplevert.

Algemene voorbeelden van continuïteitscorrecties zijn:
$P(K \geq k) = P(X > k - \frac{1}{2})$
$P(K \leq k) = P(X < k + \frac{1}{2})$
$P(K > k) = P(X > k + \frac{1}{2})$
$P(K < k) = P(X < k - \frac{1}{2})$

Overigens kan, wanneer n zeer groot is en p niet te dicht bij 0 of 1 ligt, de continuïteitscorrectie eventueel achterwege gelaten worden: de standaardafwijking $\sigma = \sqrt{np(1-p)}$ is dan relatief groot zodat de continuïteitscorrectie weinig invloed heeft op de waarde $u = \dfrac{x-\mu}{\sigma}$.

Een *derde* voorwaarde voor het mogen benaderen van een binomiale verdeling door een normale verdeling is dat de parameter n, gegeven de waarde van de parameter p, voldoende groot moet zijn. Om in te zien welke eisen we daartoe aan de parameter n in relatie tot de parameter p dienen te stellen, bedenken we het volgende.

Omdat in de binomiale verdeling met de parameters n en p de waarde van de discrete kansvariabele K niet kleiner kan zijn dan 0 (K stelt immers een *aantal* voor), betekent benadering van deze verdeling door een normale verdeling met parameters μ en σ dat de ondergrens van deze normale verdeling eveneens niet beneden 0 mag liggen. We leggen – zoals voor praktische doeleinden gebruikelijk is – deze ondergrens bij $\mu - 3\sigma$ (zie ook figuur 6.9). Immers, links daarvan ligt vrijwel geen enkele waarnemingsuitkomst. Dat betekent dus dat $\mu - 3\sigma \geq 0$ moet zijn.

Vullen we nu voor $\mu = np$ en voor $\sigma = \sqrt{np(1-p)}$ in dan ontstaat de ongelijkheid

$$np - 3\sqrt{np(1-p)} \geq 0$$

Herleiden van deze ongelijkheid leidt tot de voorwaarde dat

$$n \geq 9\frac{1-p}{p} \qquad (6.7)$$

moet zijn.
Zoals K in de binomiale verdeling niet kleiner kan zijn dan 0, zo kan K ook niet groter zijn dan n. Immers, K stelt het aantal voorkomende objecten met een bepaald kenmerk voor en dit aantal kan uiteraard niet groter zijn dan de steekproef zelf.
Op overeenkomstige wijze redenerend als hierboven is gedaan, volgt hieruit dat $\mu + 3\sigma \leq n$.
Op dezelfde manier als hierboven volgt hieruit dat n moet voldoen aan de voorwaarde

$$n \geq 9\frac{p}{1-p} \tag{6.8}$$

Dat er zowel aan formule (6.7) als aan formule (6.8) moet zijn voldaan, betekent dat n minstens gelijk moet zijn aan de grootste van de beide waarden $9\frac{1-p}{p}$ en $9\frac{p}{1-p}$.

Opdracht

Ga na dat $\dfrac{1-p}{p} > \dfrac{p}{1-p}$ wanneer $p \leq \frac{1}{2}$ en dat $\dfrac{1-p}{p} < \dfrac{p}{1-p}$ wanneer $p \geq \frac{1}{2}$.
Formule (6.7) moet dus gelden wanneer $p \leq \frac{1}{2}$ en formule (6.8) moet gelden wanneer $p \geq \frac{1}{2}$.

Voorbeeld 7

Iemand werpt 40 keer een zuivere munt op tafel. Het aantal keren dat 'kop' boven komt wordt K genoemd. Wat is de kans dat van de 40 worpen er 14 met kop boven komen?

Oplossing

Het aantal keren 'kop' is binomiaal verdeeld met parameters $n = 40$ en $p = \frac{1}{2}$. Aan voorwaarde (6.8) (en aan voorwaarde (6.7)) is voldaan, want $n = 40 > 9 \cdot \frac{0,5}{0,5} = 9$. We mogen dus de normale verdeling gebruiken als benadering van de binomiale verdeling.
Voor de parameters van deze normale verdeling moet gelden:
Normale benadering: $\mu = np = 40 \cdot 0,5 = 20$ en
$\sigma = \sqrt{np(1-p)} = \sqrt{40 \cdot 0,5 \cdot 0,5} = \sqrt{10} = 3,162$
$P_{Binomiaal}(K = 14) = P_{Normaal}(13,5 < X < 14,5) = P(X < 14,5) - P(X < 13,5)$

$P(X < 14,5) = P(U < \frac{14,5-20}{3,162}) = P(U < -1,74) = 0,0409$

$P(X < 13,5) = P(U < \frac{13,5-20}{3,162}) = P(U < -2,06) = 0,0197$

zodat $P(13,5 < X < 14,5) = 0,0409 - 0,0197 = 0,0212$
Was de vraagstelling echter: 'Gevraagd de kans op minder dan 14 keer kop, bij een worp met 40 munten', dan wordt vanwege de continuïteitscorrectie de linker overschrijdingskans berekend van 13,5.
Dus $P_B(K < 14) = P_N(X < 13,5)$.

Voorbeeld 8

In een stad met een zeer groot aantal inwoners is 60% van de kiezers vóór een zekere maatregel. Hoe groot is de kans dat een aselecte steekproef van 100 kiezers geen meerderheid oplevert voor genoemde maatregel? (Geen meerderheid betekent $K \leq 50$).

Oplossing

We passen nu de normale benadering toe om $P(K \leq 50)$ te berekenen. Ga na dat aan de voorwaarden (6.7) en (6.8) voldaan is, met $p = 0{,}6$ en $n = 100$.

$\mu = np = 100 \cdot 0{,}6 = 60$

$\sigma = \sqrt{np(1-p)} = \sqrt{100 \cdot 0{,}6 \cdot 0{,}4} = \sqrt{24}$

$P_B(K \leq 50) = P_N(X < 50{,}5) = P(U < \frac{50{,}5-60}{\sqrt{24}} = -1{,}94) = 0{,}026$

In de steekproeftheorie (hoofdstuk 7 en 8) en bij het toetsen (hoofdstuk 9) zullen we de benadering van een binomiale verdeling door een normale verdeling gebruiken.

6.5.2 Benadering van een Poisson-verdeling door een normale verdeling

Evenals de binomiale verdeling kan ook, onder bepaalde voorwaarden, de Poisson-verdeling rekentechnisch benaderd worden door een normale verdeling. Als gemiddelde voor een normale verdeling wordt het gemiddelde λ van de Poisson-verdeling genomen en als standaardafwijking $\sigma = \sqrt{\lambda}$ (dit volgens de formules voor de verwachtingswaarde en de standaardafwijking van een Poisson-verdeling, genoemd in hoofdstuk 5). Hoewel Poisson-verdelingen in principe scheef zijn (voor een kleine fractie p) is er voor grote steekproeven toch voldoende symmetrie om een Poisson-verdeling te kunnen benaderen door een normale verdeling. Voorwaarde is dat $\lambda = np > 9$.

Opdracht

Leid deze voorwaarde zelf af, op een soortgelijke manier als dat in de vorige paragraaf is ontstaan. Bedenk daarbij dat $\mu = \lambda$ en dat $\sigma = \sqrt{\lambda}$.

Ook bij de benadering van een Poisson-verdeling door een normale verdeling moet een continuïteitscorrectie worden aangebracht. De procedure is gelijk aan die is besproken bij de binomiale verdeling.

Voorbeeld 9

Stel we willen voor een Poisson-verdeling met $\lambda = 25$ de kans op een uitkomst van hoogstens 18 'successen' berekenen. Met behulp van de normale benadering kunnen we deze kans bepalen. De parameters voor de normale verdeling zijn dan: $\mu = \lambda = 25$ en $\sigma = \sqrt{\lambda} = \sqrt{25} = 5$. We beschouwen de waarde $k = 18$ als klassenmidden van de klasse met de grenzen 17,5 en 18,5. Dus om de kans op hoogstens 18 'successen' te

benaderen, bepalen we de linkeroverschrijdingskans $P_N(X < 18{,}5)$.

$$\begin{aligned} P_{Poisson}(K \leq 18) &= P_N(X < 18{,}5) \\ &= P(U < \frac{18{,}5 - 25}{5}) \\ &= P(U < -1{,}30) = P(U > 1{,}30) = 0{,}0968 \end{aligned}$$

Was de vraagstelling: 'Gevraagd de kans op minder dan 18 successen', dan moet de klasse met 18 niet meegenomen worden. We berekenen dan de linkeroverschrijdingskans $P(K < 17{,}5)$. Deze procedure volgen we ook voor rechteroverschrijdingskansen.

6.6 Negatief-exponentiële verdeling

De verdeling van het aantal gebeurtenissen per tijds- of lengte-eenheid is, zoals we gezien hebben in hoofdstuk 5, een Poisson-verdeling, als aan een aantal voorwaarden is voldaan.
Is het gemiddeld aantal gebeurtenissen per tijdseenheid gelijk aan λ, dan volgt het aantal gebeurtenissen per t tijdseenheden ook weer een Poisson-verdeling. Het gemiddeld aantal gebeurtenissen in die t tijdseenheden bedraagt dan $\mu = \lambda t$.
De tijdsduur *tussen het optreden van twee opeenvolgende gebeurtenissen* volgt nu een zogenaamde *negatief-exponentiële verdeling*.
Deze negatief-exponentiële verdeling wordt veel toegepast in 'wachttijd-problemen'. Zowel bij de tijd tussen de aankomsten bij bijvoorbeeld een loket of distributiecentrum, als voor de verdeling van de behandelingstijden aan het loket of distributiecentrum. Daarnaast speelt de negatief-exponentiële verdeling een belangrijke rol bij zogenaamde levensduurverdeling van bijvoorbeeld lampen en apparaten.
De negatief-exponentiële verdeling is een continue verdeling, dit in tegenstelling tot de Poisson-verdeling. Toch is er een nauwe verwantschap tussen de negatief-exponentiële verdeling en de Poisson-verdeling.

6.6.1 Kansdichtheid, verdelingsfunctie en eigenschappen van een negatief-exponentiele verdeling

We zullen eerst een formule afleiden voor de kansdichtheid van de negatief-exponentiële verdeling.
Het aantal gebeurtenissen K per t tijdseenheden wordt door een Poisson-verdeling met parameter $\mu = \lambda t$ beschreven:

$$P(K = k) = e^{-\lambda t}\frac{(\lambda t)^k}{k!}, \text{ met } k = 0, 1, 2, ...$$

Voor de tijdsduur T tussen twee opeenvolgende gebeurtenissen geldt:

$$P(T > t) = P(K = 0 \text{ in } t \text{ tijdseenheden}) = e^{-\lambda t}\frac{(\lambda t)^0}{0!}$$

Dus $P(T > t) = e^{-\lambda t}$ en dus geldt $P(T \leq t) = 1 - e^{-\lambda t}$
Voor deze zogenaamde *verdelingsfunctie* (= *cumulatieve kans*) van T geldt dan:

$$F(t) = P(T \leq t) = 1 - e^{-\lambda t}, \text{ voor } t \geq 0 \tag{6.9}$$

De *kansdichtheid* $f(t)$ kan afgeleid worden uit de verdelingsfunctie.

Immers: volgens de definitie van het begrip kansdichtheid is $F(t) = P(T \leq t) = \int_0^t f(t)dt$.

Wanneer we hier links en rechts van het =-teken differentiëren naar t, krijgen we volgens de wiskunde: $F'(t) = f(t)$. De kansdichtheid ontstaat dus door de verdelingsfunctie te differentiëren naar t.

$$f(t) = F'(t) = \lambda e^{-\lambda t}, \text{ voor } t \geq 0 \tag{6.10}$$

(uiteraard is $f(t) = F(t) = 0$ voor $t < 0$).

6.6.2 Verwachting en standaardafwijking van een negatief-exponentiële verdeling

Stel dat de variabele T negatief-exponentieel verdeeld is.
Voor de verwachtingswaarde μ van T geldt volgens formule (6.1):

$$\mu = E(T) = \int_0^\infty t \cdot f(t)dt = \int_0^\infty t\lambda e^{-\lambda t}dt = -\int_0^\infty t d(e^{-\lambda t}) = \left[-te^{-\lambda t}\right]_0^\infty + \int_0^\infty \lambda e^{-\lambda t}dt =$$
$$\left[0 - \frac{1}{\lambda}e^{-\lambda t}\right]_0^\infty = 0 - (-\frac{1}{\lambda}) = \frac{1}{\lambda}$$

Dus voor de verwachtingswaarde μ van de tussentijden T geldt:

$$\mu = E(T) = \frac{1}{\lambda} \tag{6.11}$$

Evenzo kan men bewijzen dat voor de standaardafwijking van T geldt:

$$\sigma = \frac{1}{\lambda} \tag{6.12}$$

In figuur 6.11 zijn voor drie waarden van μ (merk op: $\mu = \frac{1}{\lambda}$) de bijbehorende negatief-exponentiële verdelingen getekend, waarbij we de kansdichtheid kunnen schrijven als:

$$f(t) = \lambda e^{-\lambda t} = \frac{1}{\mu} \cdot e^{-\frac{t}{\mu}}$$

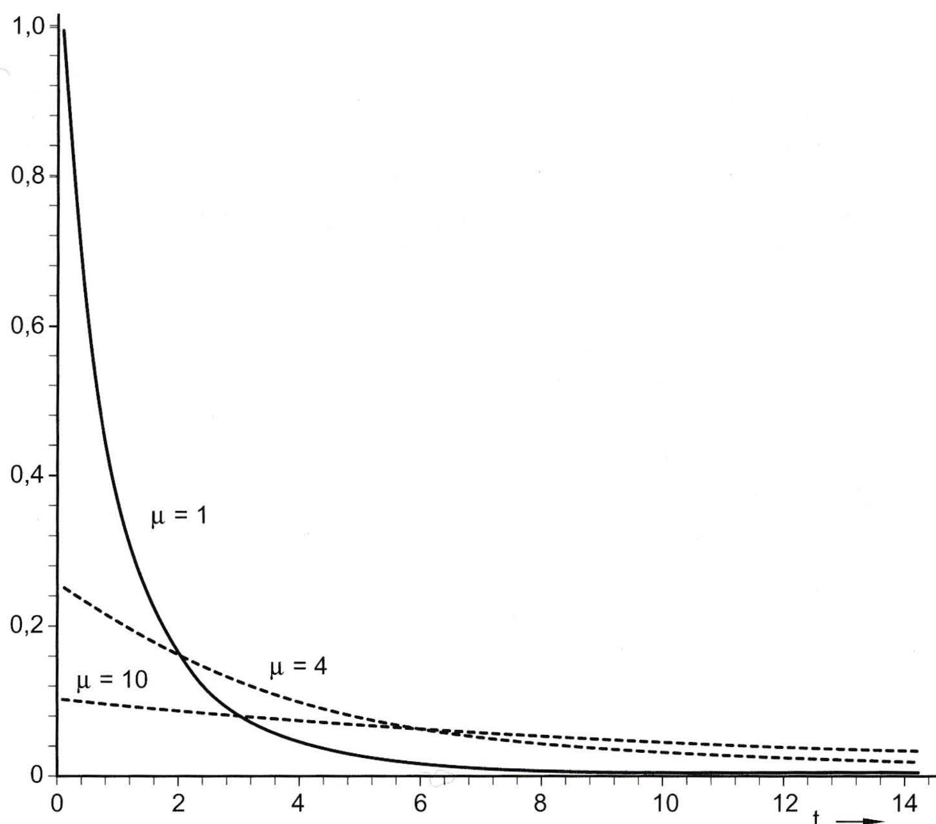


Fig. 6.13 Negatief-exponentiële verdelingen met $\mu = 1$, $\mu = 4$ en $\mu = 10$

Opmerkelijk is dat de 'top' van de verdeling steeds ligt bij $t = 0$ en niet ergens in de buurt van het gemiddelde. Zonder kennis van de verdelingsfunctie zouden we bij een gemiddelde wachttijd van bijvoorbeeld $\mu = 4$ uur een concentratie van wachttijden rond de 4 uur verwachten. Wachttijden tussen 0 en 2 uur zullen echter frequenter voorkomen dan wachttijden tussen 3 en 5 uur.

In de praktijk wordt veel gewerkt met de rechteroverschrijdingskansen $P(T \geq t)$ van de verdeling, bijvoorbeeld bij onderhoudsproblemen, levensduur, overlevingskansen enzovoorts.

$$P(T \geq t) = 1 - F(t) = e^{-\lambda t} = e^{-\frac{t}{\mu}} \tag{6.13}$$

Dit is de kans dat een tijdsduur T groter is dan, of gelijk is aan een gegeven waarde t. Deze kans is gelijk aan de kans dat gedurende een bepaalde tijd t *geen* gebeurtenis optreedt.

Voorbeeld 10

Bij een loket vervoegen zich gemiddeld 80 personen per uur. Indien deze personen volkomen toevallig en onderling onafhankelijk arriveren, zullen de tijden tussen twee willekeurige aankomsten negatief-exponentieel verdeeld zijn. Werken we niet met uren maar met minuten, dan is $\lambda = \frac{80}{60} = 1{,}33$ aankomsten per minuut en bedraagt de gemiddelde tussentijd (tijd tussen twee opeenvolgende aankomsten) $\mu = \dfrac{1}{\lambda} = \frac{1}{1{,}33} = 0{,}75$ minuten.

Beschouwen we 100 tussentijden dan verwachten we de volgende aantallen:
voor de klasse met een tussentijd 0 − 1 minuut: $100\{P(T \geq 0) - P(T \geq 1)\}$
$P(T \geq 0) = 1$
$P(T \geq 1) = e^{-\dfrac{t}{\mu}} = e^{-\frac{1}{0{,}75}} = 0{,}2636$
Dus $100\{P(T \geq 0) - P(T \geq 1)\} = 100(1 - 0{,}2636) = 73{,}64$, afgerond 74.
Voor de klasse met een tussentijd 1 − 2 minuten: $100\{P(T \geq 1) - P(T \geq 2)\}$
$P(T \geq 2) = e^{-\frac{2}{0{,}75}} = 0{,}0695$
Dus $100\{P(T \geq 1) - P(T \geq 2)\} = 100(0{,}2636 - 0{,}0695) = 19{,}41 \approx 19$ (afgerond).
Uitgewerkt voor alle klassen vinden we dan:

tussentijd min	verwachte aantallen afgerond
0 − 1	74
1 − 2	19
2 − 3	5
3 − 4	1
≥ 4	1

Voorbeeld 11

Een draad wordt gesponnen uit 1000 filamentdraden en gewikkeld op een klos. Als een van de filamentdraden breekt, stopt de opwikkelmachine automatisch. Aangenomen mag worden dat draadbreuken volkomen toevallig en onafhankelijk van elkaar optreden. Stel dat de machine gemiddeld 4 keer per uur stopt wegens een draadbreuk ($\lambda = 4$ br/uur), dan is $\mu = \frac{1}{\lambda} = \frac{1}{4}$ uur = 15 minuten. Na het repareren van een breuk en het weer aanzetten van de machine, mag men dus verwachten dat de machine gemiddeld 15 minuten zal blijven draaien voordat er ergens weer een draadbreuk optreedt. Wat kan nu gemiddeld genomen gezegd worden over de resterende tussentijd, als er sinds de laatste draadbreuk reeds 15 minuten verstreken zijn?

Oplossing

De vraagstelling suggereert een antwoord in de geest van: '*Het gaat al 15 minuten goed, dus moet er wel gauw een breuk komen*'. Niets is echter minder waar, want onafhankelijk

van de reeds verstreken tussentijd blijft de verwachting voor de tijd tot de volgende breuk gemiddeld 15 minuten.
We zullen deze interessante eigenschap van de exponentiële verdeling nu algemeen bewijzen. Stel dat T de levensduur is van een apparaat en dat T een negatief-exponentiële verdeling volgt. Wat is nu de kans dat het apparaat nog werkt op het tijdstip $t + t_0$ onder de voorwaarde dat het nog functioneert op het tijdstip t_0?
Het is duidelijk dat we hier met een voorwaardelijke kans te maken hebben:

$$P(T \geq t + t_0 | T \geq t_0)$$

Volgens de definitie van voorwaardelijke kansen (zie hoofdstuk 4: $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$) schrijven we hiervoor:

$$\frac{P(T \geq t + t_0 \text{ én } T \geq t_0)}{P(T \geq t_0)} = \frac{P(T \geq t + t_0)}{P(T \geq t_0)} = \frac{e^{-\lambda(t+t_0)}}{e^{-\lambda t_0}} = e^{-\lambda t} = P(T \geq t)$$

Conclusie: $P(T \geq t + t_0 | T \geq t_0) = P(T \geq t)$. *Dat wil zeggen de kans dat het apparaat nog werkt op het tijdstip $t + t_0$, onder de voorwaarde dat het nog functioneert op het tijdstip t_0, is onafhankelijk van t_0.* Anders gezegd: 'Op elk tijdstip t_0, waarop het apparaat nog werkt, is de kansverdeling van de (resterende) levensduur dezelfde, als de *oorspronkelijke* kansverdeling.' Er treedt geen veroudering of slijtage op: 'zolang het apparaat nog werkt, is het geheel nieuw'.
In wiskundige termen betekent dit:

$$E(T \mid T \geq t_0) = t_0 + E(T)$$

(het bewijs hiervan zullen we achterwege laten).

Opgaven

1. Stel dat X een normale verdeling volgt met $\mu = 65$ en $\sigma = 20$.
Wat is de kans dat X een waarde aanneemt:
- **a.** kleiner dan 53,2?
- **b.** groter dan 70,0?
- **c.** tussen 83,2 en 95,7?
- **d.** tussen 61,2 en 68,8?
- **e.** tussen 35,6 en 45,6?
- **f.** kleiner dan 58,6 of groter dan 70,9?

2. Een bepaald product heeft een brutogewicht van 1175 gram, met een standaardafwijking van 15 gram. De gewichten zijn normaal verdeeld. Een luchtvaartmaatschappij

accepteert dit product voor transport per vliegtuig alleen dan wanneer het minder weegt dan 1200 gram.
In hoeveel procent van de gevallen wordt het product voor transport per vliegtuig geweigerd?

3. In een jamfabriek vult men machinaal de potten jam. Wanneer het bedrijf er vrijwel zeker van wil zijn dat er minstens 250 gram in een pot zit, op welk gemiddeld gewicht moet de machine worden ingesteld, opdat de kans op een lager gewicht dan 250 gram slechts 1% bedraagt? De standaardafwijking van het vulproces bedraagt 2,5 gram.

4. Het kopergehalte van een type messing is normaal verdeeld met een gemiddelde van 70%. Bij een uitgebreid onderzoek van dit type messing vond men bij 10.000 monsters 735 monsters met een kopergehalte van meer dan 77,25%.
Bereken een schatting van de standaardafwijking van het kopergehalte in dit type messing.

5. In een fabriek maakt men ellipsvormige plaatjes, die men met behulp van schuurpoeder afslijpt tot een dikte van 72 micron. De toegestane tolerantie bedraagt $\pm$ 6 micron. Plaatjes die na het slijpen aan de tolerantie voldoen, worden met een winst van 0,10 euro per plaatjes verkocht. Te dunne plaatjes moeten worden vernietigd, hetgeen op een schade van 0,25 euro per plaatje komt. Te dikke plaatjes kunnen alsnog worden bijgeslepen, hetgeen op een extra investering komt van 2,50 euro per 100 plaatjes. Na het bijslijpen voldoen deze plaatjes aan de toleranties.
De dikte van de geproduceerde plaatjes volgt een normale verdeling met een gemiddelde dikte van 70 micron en een standaardafwijking van 4 micron.
Hoeveel winst kan de fabrikant verwachten bij een productie van 10.000 plaatjes?

6. Voor de afmeting van een bepaald product zijn tolerantiegrenzen voorgeschreven. De onderste en de bovenste waarde van de afmeting van het product zijn op grond van kwaliteitseisen opgesteld. De onderste tolerantiegrens (onderste waarde) bedraagt 16,90 mm. De afmetingen van het product hebben een normale verdeling met: $\mu = 17{,}22$ mm en $\sigma = 0{,}16$ mm. Van de geproduceerde exemplaren valt 6,28% buiten de tolerantiegrenzen.
Waar ligt de bovenste tolerantiegrens?

7. Een fabrikant heeft met zijn afnemers de volgende afspraak gemaakt. De afnemer zal uit iedere partij een steekproef nemen van 225 stuks. Ieder van deze 225 exemplaren wordt gecontroleerd op de overeengekomen kwaliteitseisen. Als er meer dan 31 exemplaren niet aan de eisen voldoen, wordt de partij teruggestuurd om vervangen te worden. Als de fabrikant een partij aflevert met 10% uitval, hoe groot is de kans dat de fabrikant de partij terug krijgt?

8. Bij gebruik van een zekere koffieautomaat worden de bekertjes, die elk maximaal 108 ml kunnen bevatten, gevuld met koffie. De hoeveelheid koffie die de automaat per gebruik levert, volgt een normale verdeling met $\mu = 100$ ml en $\sigma = 4{,}08$ ml.
Bereken de kans dat bij honderd maal gebruik van de automaat het tenminste eenmaal voorkomt dat er een bekertje overstroomt?

9. In een stad met een groot aantal inwoners is 60% vóór een zekere maatregel. Hoe groot is de kans dat een aselecte steekproef van 100 inwoners geen meerderheid oplevert voor de genoemde maatregel (de meerderheid is de helft plus één stem)?

10. Het aantal klanten dat een loketbeambte van een bioscoop per minuut bedient, is gemiddeld 2 (en Poisson-verdeeld). Hoe groot is de kans dat een bediening meer dan 10 seconden duurt?

11. Op een bepaalde plaats staat een openbare telefooncel. De gespreksduur T van een telefoongesprek is negatief-exponentieel verdeeld met een gemiddelde van 2 minuten per gesprek.

a. Hoe groot is de kans dat een gesprek meer dan 2 minuten duurt?

b. Bepaal de gemiddelde kosten per gesprek als de kosten per gesprek zijn:
- 0,1 euro bij een duur T van hoogstens 2 minuten;
- $(2t + 0{,}5)$ euro bij een duur T van meer dan 2 minuten.

12. Een randomgenerator op een zakrekenmachine produceert willekeurige getallen X tussen 0 en 1.

a. Welke verdeling volgt X, aangenomen dat de generator goed 'random' is? Geef ook de kansdichtheid van de verdeling.

b. Bepaal het gemiddelde en de standaardafwijking van X.

13. Het aantal verkeersongevallen op een druk kruispunt is Poisson-verdeeld met een gemiddelde van 15 per maand.

a. Hoe groot is de kans dat op dat kruispunt in een bepaalde maand meer dan 15 verkeersongevallen plaatsvinden?

b. Neem aan dat een maand uit 30 dagen bestaat. Wat is de kans dat tussen 2 opeenvolgende ongevallen meer dan 3 dagen ligt?

c. Gedurende een periode van 5 dagen is er geen ongeluk gebeurd op het betreffende kruispunt. Hoe groot is de kans dat gedurende de volgende 2 dagen er nog steeds geen ongeluk is gebeurd?

7 Inleiding tot de steekproeftheorie

7.1 Inleiding

Bij het nemen van steekproeven komt het vaak voor dat we te maken hebben met twee of meer kansvariabelen, die om bepaalde redenen bij elkaar moeten worden opgeteld. We denken hier bijvoorbeeld aan de totaaldikte van twee op elkaar gelaste strippen of aan de kwartaalomzetcijfers van een bepaald artikel als som van de drie opeenvolgende maandomzetcijfers van dat artikel. Ook komt het voor dat twee verschillende kansvariabelen met elkaar vergeleken moeten worden qua grootte. We kijken dan naar het verschil van de variabelen. Denk hierbij bijvoorbeeld aan de speling tussen een moer en de bijbehorende bout als het verschil tussen de (inwendige) moerdiameter en de boutdiameter. Ten slotte komt het vaak voor dat de verschillende waarden die behoren bij (een aantal exemplaren van) *dezelfde* kansvariabele moeten worden opgeteld. Denk hierbij aan de bepaling van het gemiddelde van een steekproef. Daartoe worden bijvoorbeeld n waarnemingsuitkomsten, behorend bij dezelfde variabele opgeteld, waarna de som door n gedeeld wordt.
Bij dit soort samenvoegingen van twee of meer kansvariabelen rijzen er vragen ten aanzien van de vorm van de verdeling van de samengestelde variabele en ten aanzien van het gemiddelde en de standaardafwijking van die verdeling. Het ligt voor de hand dat er voor het gemiddelde en de standaardafwijking van de som- of verschilvariabele een relatie is met het gemiddelde en de standaardafwijking van de samenstellende kansvariabelen. We zullen daarom in dit hoofdstuk de theorie van het optellen en het aftrekken van twee of meer kansvariabelen aan de orde stellen. In aansluiting daarop zullen we het gedrag van de som van meerdere (onderling onafhankelijke) kansvariabelen respectievelijk het gedrag van het gemiddelde van aselecte steekproeven uit een populatie vastleggen in de zogenaamde *centrale limietstelling*. Deze stelling vormt de basis voor de theorie, die wij de steekproeftheorie zullen noemen. In het volgende hoofdstuk zullen we die theorie daadwerkelijk gaan toepassen.

7.2 De som en het verschil van twee normaal verdeelde onderling onafhankelijke kansvariabelen

In deze paragraaf zullen we bekijken wat het betekent als twee kansvariabelen bij elkaar worden opgeteld of van elkaar worden afgetrokken.

7.2.1 De som van twee onafhankelijke normaal verdeelde kansvariabelen

Stel dat we beschikken over twee dozen A en B met daarin een groot aantal metalen strippen. De dikte X van de strippen in doos A is normaal verdeeld verondersteld met een gemiddelde $\mu_X = 1{,}6$ mm en een standaardafwijking $\sigma_X = 0{,}20$ mm. De dikte Y van de strippen in doos B is eveneens normaal verdeeld verondersteld, echter met een gemiddelde $\mu_Y = 1{,}5$ mm en een standaaardafwijking $\sigma_Y = 0{,}15$ mm. We pakken willekeurig een strip uit doos A, meten hiervan de dikte x_i, pakken vervolgens willekeurig een strip uit doos B, meten hiervan de dikte y_i, lassen de beide strippen op elkaar en meten ten slotte de dikte z_i van de 'dubbelstrip'. Deze procedure wordt herhaald totdat een van de twee dozen leeg is. Op deze wijze ontstaat een groot aantal sommen $z_i = x_i + y_i$ $(i = 1, 2, ...)$. Voor de eerste 10 dubbelstrippen krijgen we bijvoorbeeld het volgende resultaat:

i	x_i (mm)	y_i (mm)	$z_i = x_i + y_i$ (mm)
1	1,45	1,60	3,05
2	1,56	1,48	3,04
3	1,80	1,72	3,52
4	1,38	1,34	2,72
5	1,65	1,45	3,10
6	1,70	1,65	3,35
7	1,61	1,30	2,91
8	1,83	1,47	3,30
9	1,66	1,35	3,01
10	1,90	1,58	3,48

Stel dat we van de vele waarden z_i een relatieve frequentieverdeling opstellen en daar een histogram van maken. Dit histogram kan zodanig verticaal geschaald worden dat het oppervlak eronder 1 is. Dan zal – zoals we dadelijk in een stelling zullen formuleren – deze relatieve frequentieverdeling beschreven kunnen worden door een Gauss-kromme met een gemiddelde $\mu_Z = 3{,}1$ mm en een standaardafwijking $\sigma_Z = 0{,}25$ mm.
Dat de kansvariabele Z (de dikte van de dubbelstrippen) normaal verdeeld blijkt te zijn, komt doordat X en Y (de dikten van de beide afzonderlijke strippen) zelf ook normaal verdeeld zijn (een bewijs van deze stelling wordt achterwege gelaten). Zouden X en/of Y niet normaal verdeeld zijn, dan zou ook Z niet normaal verdeeld zijn.

Het gemiddelde van de somvariabele Z blijkt gelijk te zijn aan 3,1 mm. Dit is een gevolg van het feit dat $\mu_Z = \mu_X + \mu_Y = 1{,}6 + 1{,}5 = 3{,}1$.
Een dergelijke stelling geldt niet voor de standaardafwijkingen maar wel voor de varianties: $\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2 = (0{,}20)^2 + (0{,}15)^2 = 0{,}0625$, zodat $\sigma_Z = 0{,}25$. Deze stelling is echter niet algemeen geldig! De stelling geldt weliswaar ook wanneer X en Y niet normaal verdeeld zijn, maar dan wel onder de voorwaarde dat X en Y onderling onafhankelijk zijn, dus wanneer de waarden van X en Y willekeurig – onafhankelijk van elkaar – worden gekozen.[1]
Het bovenstaande is in de volgende stelling te formuleren.

Stelling 1

Als X en Y normaal verdeeld zijn, is de som $Z = X + Y$ normaal verdeeld. Het gemiddelde van Z is de som van de gemiddelden van X en Y en de variantie van Z is – mits X en Y onderling onafhankelijk zijn – gelijk is aan de som van de varianties van X en Y.
Wat betreft het gemiddelde en de variantie van de som $Z = X + Y$ geldt dus in formulevorm:

$$\mu_Z = \mu_X + \mu_Y \qquad (7.1)$$

$$\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2 \qquad (7.2)$$

Voorbeeld 1

Stel dat de op elkaar gelaste stripcombinatie (in totaal) minstens 2,8 mm dik moet zijn, hoeveel procent ervan zal dan niet aan deze eis voldoen?

Oplossing

We weten nu dat Z normaal verdeeld is met een gemiddelde $\mu_Z = 3{,}1$ mm en een standaardafwijking $\sigma_Z = 0{,}25$ mm.
We vinden dan:

$$P(Z < 2{,}8) = P\left(U < \frac{2{,}8 - 3{,}1}{0{,}25}\right) = P(U < -1{,}20) = P(U > 1{,}20) = 0{,}1151.$$

Van ongeveer 11,5% van de dubbelstrippen de zal dikte niet aan de gestelde eis voldoen.

Voorbeeld 2

Stel dat steeds twee willekeurig gekozen strippen uit doos A op elkaar gelast worden. Hoe is de dikte van de dubbelstrippen verdeeld?

[1] Als X en Y wel in zekere mate van elkaar afhankelijk zijn, zie paragraaf 10.8.

Oplossing

Er geldt nu: $Z = X + X$. Let op: dit is niet gelijk aan $2X$, hoe vreemd dat ook lijkt. Er wordt een willekeurige waarde van X bij een andere willekeurige waarde van X opgeteld. De beide waarden hoeven niet hetzelfde te zijn!
Passen we stelling 1 toe, dan blijkt dat Z normaal verdeeld is met $\mu_Z = \mu_X + \mu_X = 2\mu_X$ en $\sigma_Z^2 = \sigma_X^2 + \sigma_X^2 = 2\sigma_X^2$, zodat $\sigma_Z = \sigma_X\sqrt{2}$. Blijkbaar is $\mu_Z = 2 \cdot 1{,}6 = 3{,}2$ mm en $\sigma_Z = 0{,}20\sqrt{2} \approx 0{,}28$ mm.

Opdracht

Als een dubbelstrip ontstaat door samenvoeging van twee strippen uit doos A, waarbij net zo lang naar de tweede strip gezocht wordt tot men er een heeft gevonden die precies even dik is als de eerste strip, wat is dan het antwoord op de vraag van voorbeeld 1? Bedenk dat voor elke i geldt dat $y_i = x_i$ en dus $z_i = x_i + x_i = 2x_i$. Alle waarnemingsuitkomsten worden nu dus feitelijk met 2 vermenigvuldigd.

7.2.2 Het verschil van twee onafhankelijke normaal verdeelde kansvariabelen

Stel dat men in een ijzerhandel onder andere moeren en bouten verkoopt die in afzonderlijke bakken in voorraad worden gehouden. Bekend is dat de binnendiameter X van de moeren normaal verdeeld is met een gemiddelde $\mu_X = 5{,}8$ mm en een standaardafwijking $\sigma_X = 0{,}60$ mm. Ook de diameter Y van de bouten is normaal verdeeld, echter met een gemiddelde $\mu_Y = 5{,}0$ mm en een standaardafwijking $\sigma_Y = 0{,}45$ mm. Wanneer men bij iedere willekeurig gepakte moer een willekeurig gepakte bout voegt, kan men zich afvragen of de bout goed bij de moer past. Met andere woorden: past de bout wel bij de moer en is de speling niet te groot? We kunnen deze vraag beantwoorden als we naar het verschil van de diameter van moer en die van de bout kijken. Voor de i-de gepakte moer-boutcombinatie zal de speling, dat wil zeggen het verschil v_i tussen de moerdiameter x_i en de boutdiameter y_i gelijk zijn aan $v_i = x_i - y_i$, zodat in het algemeen geldt dat $V = X - Y$. De variabele $V = X - Y$ is te beschouwen als de som van de onderling onafhankelijke normaal verdeelde variabelen X en $-Y$, zodat $V = X + (-Y)$. Volgens stelling 1 is V dan normaal verdeeld met volgens formule (7.1) een gemiddelde $\mu_V = E(V) = E(X) + E(-Y) = E(X) - E(Y) = \mu_X - \mu_Y = 5{,}8 - 5{,}0 = 0{,}8$ mm en met volgens formule (7.2) een variantie $\sigma_V^2 = var(V) = var(X) + var(-Y) = var(X) + var(Y)$.
In dit geval geldt dus

$$\sigma_V^2 = \sigma_X^2 + \sigma_Y^2 = (0{,}60)^2 + (0{,}45)^2 = 0{,}5625 \text{ zodat } \sigma_V = 0{,}75 \text{ mm}$$

Het bovenstaande kunnen we in de volgende stelling formuleren.

Stelling 2

Het verschil $V = X - Y$ van twee kansvariabelen X en Y is normaal verdeeld, wanneer zowel X als Y normaal verdeeld is. V heeft een gemiddelde dat gelijk is aan het

verschil van de gemiddelden van X en Y en een variante die – mits X en Y onderling onafhankelijk zijn – gelijk is aan de som van de varianties van X en Y.
Voor $V = X - Y$ geldt dus in formulevorm:

$$\mu_V = \mu_X - \mu_Y \tag{7.3}$$
$$\sigma_V^2 = \sigma_X^2 + \sigma_Y^2 \tag{7.4}$$

Voorbeeld 3
Van hoeveel procent van de moer-boutcombinaties zal de speling groter zijn dan 0,5 mm?

Oplossing
Een moer-boutcombinatie zal een speling hebben van meer dan 0,5 mm wanneer er geldt: $X - Y > 0{,}5$. Met $V = X - Y$, $\mu_V = 0{,}8$ en $\sigma_V = 0{,}75$ vinden we dan:

$$\begin{aligned} P(X - Y > 0{,}5) &= P(V > 0{,}5) \\ &= P\left(U > \tfrac{0{,}5-0{,}8}{0{,}75}\right) = P(U > -0{,}40) \\ &= 1 - P(U < -0{,}40) = 1 - 0{,}3446 = 0{,}6554 \end{aligned}$$

Daarom zal van ruim 65% van de moer-boutcombinaties de speling groter zijn dan 0,5 mm.

Opdracht
Beantwoord dezelfde vraag als in voorbeeld 3 wanneer een moer-boutcombinatie ontstaat door samenvoeging van een moer en een bout, waarbij (geautomatiseerd) net zo lang naar een moer gezocht wordt tot men er een heeft gevonden waarvan de diameter precies 10% meer bedraagt dan die van de bout, zodat dan voor elke i geldt dat $x_i = 1{,}1 - y_i$ dus $z_i = (1{,}1)y_i - y_i = 0{,}1y_i$.

7.3 De som van meer dan twee onderling onafhankelijke kansvariabelen: de Centrale Limietstelling

In paragraaf 7.2 hebben we in de stellingen 1 en 2 vastgelegd dat de som respectievelijk het verschil van twee onderling onafhankelijke normaal verdeelde kansvariabelen zelf ook weer normaal verdeeld is. Het ligt voor de hand dat de stellingen 1 en 2 en de formules (7.1) t/m (7.4) kunnen worden uitgebreid wanneer meer dan twee onderling onafhankelijke kansvariabelen worden opgeteld. Dit leidt tot de volgende stelling.

Stelling 3
De som $Z = X_1 \pm X_2 \pm X_3 \pm \ldots\ldots \pm X_n$ van n normaal verdeelde kansvariabelen X_i $(i = 1, 2, 3, \ldots, n)$ bezit een normale verdeling. Het gemiddelde van Z is gelijk aan de

som van de gemiddelden van de afzonderlijke kansvariabelen. De variantie is – mits de afzonderlijke kansvariabelen paarsgewijs onafhankelijk zijn – gelijk aan de som van de varianties van de afzonderlijke kansvariabelen.
Wat betreft het gemiddelde en de variantie van $Z = X_1 \pm X_2 \pm X_3 \pm \ldots \pm X_n$, geldt dus in formulevorm:

$$\mu_Z = \mu_{X_1} + \mu_{X_2} + \mu_{X_3} + \ldots + \mu_{X_n} \tag{7.5}$$

$$\sigma_Z^2 = \sigma_{X_1}^2 + \sigma_{X_2}^2 + \sigma_{X_3}^2 + \ldots + \sigma_{X_n}^2 \tag{7.6}$$

Voorbeeld 4

Stel dat we beschikken over 10 dozen. In elke doos bevinden zich vele genummerde plaatjes. De getallen op de plaatjes bezitten per doos een normale verdeling met gemiddelde en standaardafwijking zoals weergegeven in de volgende tabel:

doos nr. i	1	2	3	4	5	6	7	8	9	10
μ_i	80	75	50	60	40	35	40	70	65	35
σ_i	2	3	4	2	5	6	5	3	4	5

Uit ieder van de tien dozen nemen we een plaatje, noteren het getal dat er op staat en bepalen de som S van deze 10 getallen. Hoe groot is de kans dat S groter zal zijn dan 580?

Oplossing

Omdat de getallen op de plaatjes per doos normaal verdeeld zijn, zal volgens stelling 3 ook de som S normaal verdeeld zijn en wel volgens formule (7.5) met gemiddelde $\mu_S = 80 + 75 + 50 + 60 + 40 + 35 + 40 + 70 + 65 + 35 = 550$ en volgens formule (7.6) met variantie $\sigma_S^2 = 2^2 + 3^2 + 4^2 + 2^2 + 5^2 + 6^2 + 5^2 + 3^2 + 4^2 + 5^2 = 169$ dus met standaardafwijking $\sigma_S = 13$.
We vinden dan:

$$P(S > 580) = P\left(U > \frac{580 - 550}{13}\right) = P(U > 2{,}31) = 0{,}0104$$

Dat de in stelling 3 bedoelde som $Z = X_1 \pm X_2 \pm X_3 \pm \ldots \pm X_n$ normaal verdeeld is, komt door het feit dat elk van de variabelen X_i $(i = 1, 2, 3, \ldots, n)$ zelf ook normaal verdeeld is. Echter, ook wanneer de samenstellende variabelen zelf geen van alle of niet alle normaal verdeeld zijn, zal toch hun som, althans bij goede benadering, wel normaal verdeeld zijn. Deze benadering zal des te beter zijn naarmate het aantal samenstellende variabelen (n) groter is. Of anders gezegd: hoe groter het aantal samenstellende variabelen is, hoe minder noodzakelijk het is dat deze normaal verdeeld zijn om hun som wel normaal verdeeld te doen zijn.

Wat het speciale geval $Z = X_1 \pm X_2 \pm X_3 \pm \ldots \pm X_n = \sum_{i=1}^{n} X_i$ betreft, is dit opmerkelijke feit vastgelegd in stelling 4, de *Centrale Limietstelling van Laplace*.

Stelling 4

De som $Z = \sum_{i=1}^{n} X_i$ van n kansvariabelen X_i $(i = 1, 2, 3, \ldots, n)$ die al of niet normaal verdeeld zijn, is bij goede benadering normaal verdeeld – en des te beter naarmate n groter is. Het gemiddelde van Z is gelijk aan de som van de gemiddelden van de afzonderlijke kansvariabelen. De variantie van Z is – mits de afzonderlijke kansvariabelen paarsgewijs onafhankelijk zijn – gelijk aan de som van de varianties van de afzonderlijke kansvariabelen.

Wat betreft het gemiddelde en de variantie van de in stelling 4 bedoelde som $Z = \sum_{i=1}^{n} X_i$ geldt in formulevorm:

$$\mu_Z = \sum_{i=1}^{n} \mu_i \qquad (7.7)$$

$$\sigma_Z^2 = \sum_{i=1}^{n} \sigma_i^2 \qquad (7.8)$$

De formules (7.7) en (7.8) gelden zowel in het geval dat de X_i $(i = 1, 2, 3, \ldots, n)$ alle normaal verdeeld zijn als in het geval dat zij geen van alle of niet alle normaal verdeeld zijn. Bedenk wel dat de gesommeerde variabelen onderling onafhankelijk moeten zijn.

Voorbeeld 5

Bij de fabricage van een bepaald soort product worden om een metalen staaf, die op een voetstukje is gemonteerd, afwisselend aluminium ringen en aluminium platen geschoven, eerst een ring, dan een plaat, vervolgens weer een ring, enzovoorts. Na de laatste (de 16de) ring wordt het geheel afgesloten met een schroefdop, waarna het nog boven de dop uitstekende stuk van de staaf wordt afgezaagd. Bereken het percentage producten waarbij het staafrestant langer is dan 8 mm wanneer we beschikken over de volgende gegevens:

- de lengte van de metalen staven is normaal verdeeld met een gemiddelde van 130 mm en een variantie van 0,21 mm^2;
- de ringen hebben een dikte met een gemiddelde van 5 mm en een standaardafwijking van 0,2 mm;
- de platen hebben een dikte met een gemiddelde van 2 mm en een standaardafwijking van 0,3 mm;

- de afsluitdoppen hebben een dikte met een gemiddelde van 10 mm en een standaardafwijking van 0,6 mm;
- de dikte van de schroefdoppen is wel, maar de dikte van de ringen en de platen is niet normaal verdeeld;
- alle variabelen zijn onderling onafhankelijk.

Oplossing

We noemen de dikte van de ringen (16 stuks), de platen (15 stuks) en de schroefdop achtereenvolgens R, P en D, de lengte van de staven en staafrestanten achtereenvolgens S en V en de hoogte van de ring-plaat-dop combinatie C. Volgens de verstrekte gegevens geldt dan:
$C = R + P + R + ... + R + P + R + D$ en volgens stelling 4 is C normaal verdeeld (ook al zijn R en P dat niet). Verder geldt er dat $V = S - C$ normaal verdeeld is omdat zowel S als C dat is. We vinden dan:

volgens formule (7.7): $\mu_C = \mu_R + \mu_P + \mu_R + ... + \mu_R + \mu_P + \mu_R + \mu_D = 16\mu_R + 15\mu_P + \mu_D = 80 + 30 + 10 = 120.$

volgens formule (7.8): $\sigma_C^2 = \sigma_R^2 + \sigma_P^2 + \sigma_R^2 + ... + \sigma_R^2 + \sigma_P^2 + \sigma_R^2 + \sigma_D^2 = 16 \cdot \sigma_R^2 + 15 \cdot \sigma_P^2 + \sigma_D^2 = 0{,}64 + 1{,}35 + 0{,}36 = 2{,}35.$

volgens (7.3): $\mu_V = \mu_S - \mu_C = 130 - 120 = 10.$

volgens (7.4): $\sigma_V^2 = \sigma_S^2 + \sigma_C^2 = 0{,}21 + 2{,}35 = 2{,}56,$
dus $\sigma_V = 1{,}6.$

Hieruit volgt:

$$P(V > 8) = P\left(U > \frac{8-10}{1{,}6}\right) = P(U > -1{,}25)$$
$$= 1 - P(U < -1{,}25) = 1 - P(U > 1{,}25) = 1 - 0{,}1056 = 0{,}8944$$

Daarom zal bij bijna 90% van de producten het staafrestant langer zijn dan 8 mm.

We zullen de Centrale Limietstelling in dit boek niet bewijzen. Overigens wordt de stelling vaak iets anders geformuleerd dan wij hierboven in stelling 4 hebben gedaan. Stelling 4 heeft betrekking op de som van een aantal paarsgewijs onafhankelijke al of niet normaal verdeelde kansvariabelen met verschillende gemiddelden en verschillende standaardafwijkingen. In zijn andere formulering heeft de Centrale Limietstelling betrekking op het gemiddelde van een aantal paarsgewijs onafhankelijke al of niet normaal verdeelde kansvariabelen met *hetzelfde gemiddelde en dezelfde standaardafwijking*. In paragraaf 7.4 (stelling 5) komen we hierop terug.

Men zou zich bij stelling 4 kunnen afvragen hoe groot de daarin genoemde n moet zijn om te kunnen zeggen dat de som van n niet-normaal verdeelde kansvariabelen toch een normale verdeling bezit. Het antwoord op deze vraag hangt niet alleen af van de mate waarin de diverse afzonderlijke verdelingen van een normale verdeling afwijken, maar ook van de grootte van de verschillen tussen hun gemiddelden dan wel standaardafwijkingen. In het geval dat geen van de n variabelen normaal verdeeld is, hanteert men in de praktijk vaak de vuistregel dat n minstens 25 moet zijn om te kunnen zeggen dat de som van die variabelen normaal verdeeld is.
De Centrale Limietstelling is een van de belangrijkste stellingen uit de statistiek en wordt vaak toegepast in de steekproeftheorie.

7.4 Het gemiddelde van een aselecte steekproef

Wanneer de in stelling 4 bedoelde kansvariabelen alle hetzelfde gemiddelde μ en dezelfde standaardafwijking σ bezitten, gaat stelling 4 over in de volgende stelling.

Stelling 5
De som $Z = \sum_{i=1}^{n} X_i$ van n kansvariabelen X_i, die voor elke i ($i = 1, 2, 3, ...n$) al of niet normaal verdeeld zijn met hetzelfde gemiddelde μ en dezelfde standaardafwijking σ, bezit bij benadering een normale verdeling waarvan het gemiddelde gelijk is aan:

$$\mu_Z = n \cdot \mu \tag{7.9}$$

Voor grote waarden van n is deze benadering beter dan voor kleine waarden van n.
De variantie van Z is - mits de kansvariabelen paarsgewijs onafhankelijk zijn - gelijk aan

$$\sigma_Z^2 = n \cdot \sigma^2 \tag{7.10}$$

Voorbeeld 6
In een magazijn dat een hoogte heeft van 312 cm, worden metalen schijven opgeslagen. Men heeft de gewoonte om stapels te maken van 25 schijven hoog. De dikte D van de schijven heeft een gemiddelde $\mu_D = 12$ cm en een standaardafwijking $\sigma_D = 2$ cm.
Bij hoeveel procent van de stapels zal het niet lukken deze compleet te maken?

Oplossing
De hoogte H van stapels van 25 schijven is volgens stelling 5 normaal verdeeld, ongeacht of de dikte D van de schijven zelf dat wel of niet is. Volgens formule (7.9) heeft H een gemiddelde $\mu_H = 25 \cdot \mu_D = 25 \cdot 12 = 300$ cm en volgens formule (7.10) heeft H een variantie $\sigma_H^2 = 25 \cdot \sigma_D^2 = 25 \cdot 2^2 = 100$, dus een standaardafwijking $\sigma_H = 10$ cm.

De gevraagde kans berekenen we nu als volgt:

$$P(H > 312) = P\left(U > \frac{312 - 300}{10}\right) = P(U > 1{,}20) = 0{,}1151$$

Bij ongeveer 11,5% van de stapels zal het niet lukken deze compleet te maken.

Opdracht

Beredeneer waarom het in het laatste voorbeeld onjuist is te stellen dat $H = 25 \cdot D$. Ga na ook na welk gevolg deze onjuiste veronderstelling heeft voor het gemiddelde en de standaardafwijking van H.

Stel nu dat we beschikken over een al of niet normaal verdeelde populatie met gemiddelde μ en standaardafwijking σ. Uit stelling 5 volgt dat de som $z = \sum_{i=1}^{n} x_i$ van de n paarsgewijs onafhankelijke waarnemingsuitkomsten van een aselecte steekproef uit deze populatie beschouwd kan worden als een waarde van de normaal verdeelde kansvariabele Z met gemiddelde $\mu_Z = n\mu$ en variantie $\sigma_Z^2 = n\sigma^2$, dus standaardafwijking $\sigma_Z = \sigma\sqrt{n}$.

Het gemiddelde $\bar{x} = \frac{z}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$ van een aselecte steekproef van n stuks uit een al of niet normaal verdeelde populatie met gemiddelde μ en standaardafwijking σ kan daarom beschouwd kan worden als een waarde van de normaal verdeelde kansvariabele $\bar{X} = \frac{Z}{n}$ met gemiddelde $\mu_{\bar{X}} = \frac{\mu_Z}{n} = \frac{n\mu}{n} = \mu$ en standaardafwijking $\sigma_{\bar{X}} = \frac{\sigma_Z}{n} = \frac{\sigma\sqrt{n}}{n} = \frac{\sigma}{\sqrt{n}}$.
Dit leidt tot de volgende stelling.

Stelling 6

Stel we nemen een aselecte steekproef van n stuks uit een (al of niet) normaal verdeelde populatie met gemiddelde μ en standaardafwijking σ. Stel ook dat de n waarnemingsuitkomsten x_i $(i = 1, 2, 3, ..., n)$ van de steekproef paarsgewijs onafhankelijk zijn. Het gemiddelde $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ is dan een waarde van de kansvariabele $\bar{X}$, die (althans bij benadering en des te beter naarmate n groter is) normaal verdeeld is met gemiddelde

$$\mu_{\bar{X}} = \mu \tag{7.11}$$

en met standaardafwijking

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \tag{7.12}$$

Voorbeeld 7

Door een numeriek bestuurd meetapparaat worden per uur duizenden glazen buisjes op hun inwendige diameter gesorteerd in de kwaliteitsklassen 'large', 'medium' en 'small'. In de kwaliteitsklasse 'medium' dient de gemiddelde diameter 10 mm te bedragen met een standaardafwijking van 0,05 mm.
Elk kwartier neemt de automaat uit ieder van de drie kwaliteitsklassen een steekproef van 16 buisjes, meet hiervan de inwendige diameter en berekent het gemiddelde van de 16 waarnemingsuitkomsten. Wanneer dit gemiddelde voor de steekproef uit de klasse 'medium' minder dan 9,98 mm of meer dan 10,02 mm bedraagt, wordt het sorteerproces gestopt. Hoe groot is de kans dat dit (ten onrechte) gebeurt, wanneer toch aan de voor de klasse 'medium' gestelde eis is voldaan?

Oplossing

We nemen een steekproef van 16 stuks uit een populatie met een gemiddelde $\mu = 10$ en een standaardafwijking $\sigma = 0{,}05$ en bepalen daarvan het gemiddelde. Voor het beantwoorden van de gestelde vraag dienen we te berekenen hoe groot de kans is dat het steekproefgemiddelde kleiner is dan 9,98 of groter is dan 10,02. Volgens stelling 5 zijn de gemiddelden van 16 waarnemingsuitkomsten uit een populatie met een gemiddelde $\mu = 10$ mm en een standaardafwijking $\sigma = 0{,}05$ mm normaal verdeeld. Volgens formule (7.11) hebben die steekproefgemiddelden een gemiddelde $\mu_{\bar{X}} = \mu = 10$ mm. Volgens formule (7.13) is de standaardafwijking van de steekproefgemiddelden $\sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{0{,}05}{4} = 0{,}0125$ mm. We vinden voor deze kans:

$$\begin{aligned}
P\left(\bar{X} < 9{,}98 \text{ of } \bar{X} > 10{,}02\right) &= P(\bar{X} < 9{,}98) + P(\bar{X} > 10{,}02) \\
&= P\left(U < \frac{9{,}98 - 10}{0{,}0125}\right) \\
&\quad + P\left(U > \frac{10{,}02 - 10}{0{,}0125}\right) \\
&= P(U < -1{,}60) + P(U > 1{,}60) \\
&= 2 \cdot P(U > 1{,}60) = 2 \cdot 0{,}0548 = 0{,}1096
\end{aligned}$$

Stelling 5 is in feite een bijzonder geval van de Centrale Limietstelling zoals geformuleerd in stelling 4 en is eveneens van groot belang voor de steekproeftheorie.
De in stelling 5 genoemde voorwaarde dat de waarnemingsuitkomsten in de steekproef paarsgewijs onafhankelijk moeten zijn, betekent in feite dat de steekproef met teruglegging genomen moet worden. Ieder exemplaar dat in de steekproef wordt opgenomen, zou dus eerst moeten worden teruggelegd alvorens het volgende exemplaar gepakt kan worden, waarbij elke trekking aselect (willekeurig) dient te geschieden. In de praktijk wordt hieraan meestal niet voldaan. Er wordt immers meestal zonder teruglegging getrokken. Toch kan in de meeste gevallen formule (7.12) wel degelijk gebruikt worden. Dit komt doordat

de steekproef meestal klein is ten opzichte van de populatie. Daardoor zal, ook al is de steekproef zonder teruglegging, de verhouding van het aantal elementen met een bepaald kenmerk en de populatiegrootte nauwelijks veranderen. Indien de steekproef (zonder teruglegging) relatief groot is ten opzichte van de populatie, moet een correctiefactor worden aangebracht:
Aangetoond kan worden dat in het geval van een aselecte steekproef zonder teruglegging de essentie van stelling 5 ongewijzigd blijft, echter formule (7.12) moet dan vervangen worden door

$$\sigma_{\bar{X}} = \sqrt{\frac{N-n}{N-1}} \cdot \frac{\sigma}{\sqrt{n}} \tag{7.13}$$

waarin N de populatiegrootte is.
Uit formule (7.13) blijkt dat, naarmate de populatieomvang N groter is ten opzichte van de steekproefgrootte n, de factor $\sqrt{\frac{N-n}{N-1}}$ - steeds meer de waarde 1 benadert, zodat de waarde van de standaardafwijking $\sigma_{\bar{X}}$ volgens formule (7.13) steeds meer de waarde van de standaardafwijking $\sigma_{\bar{X}}$ volgens formule (7.12) benadert.
Uit het voorafgaande volgt dat voor aselecte steekproeven zonder teruglegging formule (7.12) bij goede benadering bruikbaar is, mits de steekproeven niet al te groot zijn ten opzichte van de populatie. In de praktijk legt men als vuistregel de grens bij een factor 10: Gebruik (7.12) wanneer $N \geq 10n$ en gebruik (7.13) wanneer $N < 10n$.

Opgaven

1. Een bedrijfskundige afdeling van een fabriek heeft de organisatie van het werk zodanig ingericht dat een operator twee machines tegelijk kan bedienen. Van machine A zijn de bedieningstijden normaal verdeeld met een gemiddelde van 200 seconden en een standaardafwijking van 24 seconden. Ook van machine B zijn de bedieningstijden normaal verdeeld; het gemiddelde bedraagt 324 seconden en de standaardafwijking 32 seconden. Gezien de organisatorische opstelling is het niet gewenst dat de operator aan de bediening van beide machines samen meer dan 10 minuten besteedt. Hoe groot is de kans dat dit toch zal gebeuren?

2. Een automaat monteert ronde pennetjes in cilindervormige buisjes.
Er wordt telkens aan een willekeurig buisje een willekeurig pennetje toegevoegd. Wanneer de samenstelling niet past, wordt zowel het buisje als het pennetje weer in het aanvoercircuit opgenomen.
De diameters van de beide onderdelen zijn normaal verdeeld met voor de buisjes een gemiddelde van 11,2 mm en een standaardafwijking van 0,50 mm en voor de pennetjes een gemiddelde van 10,2 mm en een standaardafwijking van 0,375 mm.
Hoe groot is de kans dat een buisje en het bijgevoegde pennetje weer in het aanvoercircuit worden opgenomen?

3. Vier atleten A, B, C en D vormen samen een estafetteteam. Gedurende een langere periode werden hun trainingstijden gemeten over de afstand 400 m. Deze bleken normaal verdeeld te zijn met (in seconden) gemiddelden $\mu_A = 48{,}4$; $\mu_B = 50{,}6$; $\mu_C = 49{,}8$ en $\mu_D = 51{,}2$ en standaardafwijkingen $\sigma_A = 1{,}2$; $\sigma_B = 2{,}1$; $\sigma_C = 2{,}4$ en $\sigma_D = 2{,}8$. Tijdens een bepaald sportevenement zal door het viertal worden getracht het bestaande baanrecord over de 4×400 m (zijnde 190,5 seconden) te breken.

a. Hoe groot is de kans dat de recordpoging zal slagen?
b. Hoe groot is de kans dat B een betere tijd maakt dan A?
c. Hoe groot is de kans dat D de snelste tijd maakt van de vier?

4. Een bepaald soort flessen is tijdens transport aan breuk onderhevig. Wanneer een exemplaar met een breeksterkte B tijdens transport een kracht K ondervindt die groter is dan B, breekt het exemplaar. Is K kleiner dan B, dan breekt het exemplaar niet. Over een lange tijdsperiode genomen bleek bij 3,01% breuk te zijn opgetreden. Uit gedane onderzoekingen bleek in dezelfde periode de breeksterkte van de getransporteerde flessen normaal verdeeld te zijn met een gemiddelde van 200 kg en een standaardafwijking van 40 kg. Het management vond het optredende percentage breuk te hoog en besloot daarom door gewijzigde samenstelling van het glas de breeksterkte ervan op te voeren. Nadat dit was gebeurd, bleek uit nieuwe proeven de gemiddelde breeksterkte 231,5 kg te zijn geworden. De vorm van de verdeling van de breeksterkten en de standaardafwijking bleken geen wijziging te hebben ondergaan, terwijl het breukpercentage was teruggebracht tot 0,6%.

a. Als mag worden aangenomen dat de tijdens transport optredende breekkrachten normaal verdeeld zijn en onafhankelijk zijn van de breeksterkten van de flessen, hoe groot is dan het gemiddelde en de standaardafwijking van deze breekkrachten?
b. Hoe groot moet dan bij gelijkblijvende standaardafwijking de gemiddelde breeksterkte worden wanneer men bereid is 1,1% breuk te accepteren ?

5. Het ogenaantal K van een dobbelsteen is een discrete kansvariabele.

a. Teken de grafiek van de kansfunctie van K en bereken μ_K en σ_K.
b. Men werpt met twee dobbelstenen en bepaalt de som L van de ogenaantallen. Merk op dat deze som niet te schrijven is als $2 \cdot L$ (dit is het ogenaantal vermenigvuldigen met 2), maar wel als $L + L$. Bepaal de kansfunctie van L, teken de grafiek en bepaal μ_L en σ_L.
c. Men werpt met drie dobbelstenen en bepaalt de som M van de ogenaantallen. Bepaal de kansfunctie van M, teken de grafiek en bepaal μ_M en σ_M.
d. Vergelijk de grafieken van a, b en c en verklaar aan met behulp van de Centrale Limietstelling het resultaat.

6. Een firma wil voor een bepaald soort orders de doorlooptijd vanaf besteldatum tot aflevering uit de expeditieloods onderzoeken. Er zijn vier achtereenvolgende stadia waarvoor frequentieverdelingen van de (bij benadering normaal verdeelde) doorlooptijden zijn gemaakt.

Voor de gemiddelden en de standaardafwijkingen van de vier stadia vond men de volgende schattingen (in dagen):

I. Vanaf besteldatum tot aanvang fabricage): $\mu_1 = 14$ en $\sigma_1 = 2{,}0$;
II. Vanaf begin tot eind fabricage: $\mu_2 = 6$ en $\sigma_2 = 1{,}6$;
III. Opslag fabrieksmagazijn gereed product: $\mu_3 = 12$ en $\sigma_3 = 1{,}0$;
IV. Opslag in loods expeditiecentrum: $\mu_4 = 18$ en $\sigma_4 = 1{,}2$.

a. Hoeveel bedraagt het gemiddelde en de standaardafwijking van de totale doorlooptijd, aannemende dat de doorlooptijden in de afzonderlijke stadia paarsgewijs onafhankelijk zijn?
b. Tussen welke waarden zal de totale doorlooptijd in 95% van de gevallen spreiden, dat wil zeggen buiten welk interval ligt slechts 2,5% kortere doorlooptijden en 2,5% langere doorlooptijden?
c. Hoeveel bedraagt de totale doorlooptijd wanneer men een bepaalde order in elk stadium zoveel mogelijk bespoedigt, dat wil zeggen de doorlooptijd in elk stadium zodanig kiest dat deze in slechts 2,5% van de gevallen korter is?
d. Idem wanneer een order over elk stadium zo lang mogelijk doet (2,5% van de doorlooptijden duurt nog langer)?

7. Een fabriek van bakkersartikelen vervaardigt onder andere speculaasjes. De gewichten van de speculaasjes zijn scheef verdeeld met een gemiddelde van 3 gram en een standaardafwijking van 0,2 gram. De kartonnen doosjes waarin 50 speculaasjes worden verpakt, wegen gemiddeld 15 gram met een standaardafwijking van 0,5 gram (normaal verdeeld).
Hoeveel procent van de gevulde doosjes zal minder dan 162 gram wegen wanneer alle variabelen onderling onafhankelijk zijn?

8. Een levensmiddelenconcern beschikt voor het vullen van pakken koffie over een automatische vulmachine die staat afgesteld op een netto vulgewicht van 250 gram en die werkt met een standaardafwijking van 12 gram. Bekend is dat de gewichten van de lege pakjes normaal verdeeld zijn met een gemiddelde van 20 gram en een standaardafwijking van 5 gram.
Voor de verzending naar de supermarktfilialen worden 24 pakken koffie in een doos gedaan waarvan het (lege) gewicht normaal verdeeld is met een gemiddelde van 520 gram en een standaardafwijking van 13 gram.
a. Wat is het gemiddelde en de standaardafwijking van een met 24 pakken koffie gevulde verzenddoos?
b. Hoe groot is de kans dat twee dozen met elk 24 pakken koffie meer dan 100 gram in gewicht verschillen?

Een keuringsinstantie heeft bezwaar tegen de grote spreiding waarmee de vulautomaat werkt en bepaalt dat het gewicht van een doos met 24 pakken koffie slechts in 2,28% van de gevallen groter mag zijn dan 7074 gram.

c. Tot welk bedrag moet de standaardafwijking van de vulmachine worden teruggebracht om aan deze eis te kunnen voldoen?

9. Een supermarkt heeft voor een groot aantal artikelen laten nagaan hoe groot in de periode januari 2000 t/m december 2000 de wekelijkse omzet was.
Voor doosjes dadels van een bepaald merk vond men de volgende frequentieverdeling:

verkocht aantal doosjes dadels per week	aantal weken
153 t/m 157	2
158 t/m 162	4
163 t/m 167	7
168 t/m 172	13
173 t/m 177	12
178 t/m 182	7
183 t/m 187	3
188 t/m 192	2

De bedrijfsleider heeft de gewoonte om elke vier weken de winkelvoorraad (inclusief de magazijnvoorraad) voor wat betreft het bewuste artikel door een bestelling bij de importeur aan te vullen, de importeur kan dan prompt uit voorraad leveren.
Op een bepaalde besteldatum blijken er nog 250 doosjes dadels in voorraad te zijn. Hoeveel doosjes moet de bedrijfsleider op dat moment bestellen wanneer hij slechts 5% risico wil lopen de komende periode van vier weken buiten voorraad (stock out) te geraken? (Aangenomen mag worden dat vier opeenvolgende weekomzetcijfers paarsgewijs onafhankelijk zijn.)

10. Bij de fabricage van een bepaald onderdeel van wasmachines worden beurtelings vier aluminium strippen en vier koperen plaatjes op elkaar gelast. Van de dikte van de aluminium strippen is bekend dat deze normaal verdeeld is met een gemiddelde van 2,0 mm en een standaardafwijking van 0,3 mm. Voor de totale dikte van de gelaste samenstellingen geldt de technische tolerantie 21,43 $\pm$ 1,30 mm.
Van een zeer grote partij gelaste samenstellingen is de dikte normaal verdeeld met een gemiddelde van 21,60 mm. De dikte van de lasnaden is verwaarloosbaar klein.
Van deze partij is 7,08% van de samenstellingen te dun.

a. Welke standaardafwijking bezit de dikte van de gelaste samenstellingen in de partij?

b. Hoeveel procent van de gelaste samenstellingen is te dik?

c. Hoeveel procent uitval bevat de partij gelaste samenstellingen?

d. Hoe groot is het gemiddelde en de standaardafwijking van de dikte van de afzonderlijke koperen plaatjes?

e. Als de dikte van de koperen plaatjes normaal verdeeld is en moet voldoen aan de tolerantie 3,492 ± 0,804 mm, hoeveel procent van de koperen plaatjes voldoet dan niet aan deze eis?

11. De levensduur van een bepaald soort bioscooplampje is normaal verdeeld met een gemiddelde van 3250 branduren en een standaardafwijking van 400 branduren.
De bioscoopexploitanten eisen dat de lampjes minstens 3000 uur achtereen zullen branden.

a. Hoeveel procent van de door de fabrikant afgeleverde lampjes zal naar verwachting niet aan deze eis voldoen?

b. Op welke gemiddelde levensduur zal de fabrikant zijn productieproces moeten inrichten opdat behoudens 2,5% uitval aan de eis van de bioscoopexploitanten wordt voldaan?
Deze reorganisatie van het productieproces kost veel geld. De fabrikant overweegt dit te vermijden door aan de bioscoopexploitanten voor te stellen voortaan pakketjes van vier lampjes af te leveren die een gezamenlijke minimale levensduur hebben van 4 × 3000 = 12.000 branduren.

c. Hoeveel procent van de door de fabrikant af te leveren pakketjes zou niet aan deze eis voldoen?

d. Op welke gemiddelde levensduur zou de fabrikant zijn productieproces moeten inrichten opdat behoudens 2,5% uitval aan deze eis zou worden voldaan?
Uit het voorafgaande blijkt dat, om aan de eis van de bioscoopexploitanten te kunnen voldoen, het procesgemiddelde bij verkoop van pakketjes met 4 lampjes tegelijk in mindere mate verschoven behoeft te worden dan bij verkoop van individuele lampjes.

e. Bereken hoeveel lampjes per pakketje de fabrikant zou moeten verkopen opdat behoudens 0,62% uitval wordt voldaan aan de eis van de bioscoopexploitanten zonder dat het procesgemiddelde behoeft te worden verschoven.

12. Een grootwinkelbedrijf in sportartikelen berekent elke week de gemiddelde omzet van haar 25 (als even groot te beschouwen) filialen in het land. Over een heel jaar (50 weken) berekend bleken deze weekgemiddelden normaal verdeeld te zijn met een gemiddelde van 94380 euro en een standaardafwijking van 4000 euro.
Wanneer de omzet per week per filiaal geacht mag worden normaal verdeeld te zijn, hoe groot is dan de kans dat in een willekeurig filiaal een weekomzet van meer dan 120.000 euro wordt gemaakt?

13. Een bepaald type TL-buizen van de firma P heeft een gemiddelde levensduur van 14.000 uur en een standaardafwijking van 2000 uur.

Firma Q vervaardigt hetzelfde type TL-buizen met een gemiddelde levensduur van 12.000 uur en een standaardafwijking van 1000 uur.
De KEMA neemt regelmatig steekproeven van 125 stuks uit de productie van beide firma's en bepaalt voor elke steekproef onder andere de levensduur van de TL-buizen.
Hoe groot is de kans dat wat de levensduur van de TL-buizen betreft de beide steekproefgemiddelden meer dan 1800 uur verschillen?

14. De Quality Officer van een steenfabriek laat wekelijks steekproeven van 100 stuks nemen uit de magazijnvoorraad trottoirtegels. De Production Manager van het bedrijf laat dagelijks uit de productiestroom van de trottoirtegels steekproeven van 25 stuks nemen. Uit QUINFOST, het Quality Information System van het bedrijf, blijkt wat de dikte van de tegels betreft onder andere het volgende:
- 17,11 % van de steekproeven van het Quality Office hebben een gemiddelde groter dan 25,19 mm;
- 2,87% van de steekproeven van de Production Unit hebben een gemiddelde groter dan 25,76 mm.

Hoe groot is het gemiddelde en de standaardafwijking van de dikte van de trottoirtegels?

15. In een bak zitten 200 messing schroeven en 400 aluminium schroeven. De lengte van de messing schroeven is normaal verdeeld met $\mu_M = 4{,}5$ cm en $\sigma_M = 0{,}15$ cm. De lengte van de aluminium schroeven is normaal verdeeld met $\mu_A = 4{,}3$ cm en $\sigma_A = 0{,}2$ cm.

a. Men pakt aselect een aluminium schroef en een messing schroef. Hoe groot is de kans dat de aluminiumschroef langer is dan de messingschroef?

b. Men pakt 25 aluminum schroeven en bepaalt het gemiddelde van de (25) lengtes. Hoe groot is de kans dat dit gemiddelde minder dan 1 mm verschilt van 4,3 cm?

16. Bedrijf A verkoopt tomatenzaad, dat rijpe tomaten oplevert in gemiddeld 54 dagen met een standaardafwijking van 6 dagen. Bedrijf B verkoopt tomatenzaad, dat rijpe tomaten oplevert in gemiddeld 60 dagen met een standaardafwijking van 8 dagen.

a. Hoe groot is de kans dat een tomaat van zaad A meer tijd nodig heeft om tot rijping te komen dan een tomaat van zaad B?

b. Een tomatenkweker plant 400 zaden van beide firma's. Van alle tomaten, afkomstig van A-zaad wordt het gemiddelde aantal dagen van planten tot rijpheid berekend ($M_A = m_A$). Van de 400 tomaten van B-zaad idem dito ($M_B = m_B$).
Hoe groot is de kans dat M_A meer dan 2 dagen van M_B verschilt?

8 Schatten

8.1 Inleiding

We vervolgen nu met het tweede gedeelte van de steekproeftheorie. Nu zullen we daadwerkelijk steekproeven uit populaties nemen om de karakteristieke grootheden van de populatie te weten te komen. We hebben het dan over het *schatten* van populatieparameters.

8.2 Het schatten van populatieparameters

In de laatste paragraaf van het voorgaande hoofdstuk werd feitelijk de basis gelegd voor de zogenaamde steekproeftheorie. We zullen daar nu dieper op ingaan. Een steekproef uit een populatie wordt meestal genomen om een schatting te maken van de parameters van de populatie. Zo kan het gemiddelde μ van een populatie geschat worden door het steekproefgemiddelde $\bar{x}$.
De standaardafwijking σ van de populatie kan geschat worden door de standaardafwijking s van de steekproef. Zoals in hoofdstuk 3 al uiteengezet is blijkt $s = \sqrt{\dfrac{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$ een betere schatter van σ te zijn dan $s = \sqrt{\dfrac{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}{n}}$.
De fractie p van een populatie wordt geschat door de fractie $\hat{p}$ van de steekproef (merk op dat we voor de fractie p van de populatie bij wijze van uitzondering geen Griekse letter schrijven).
Wanneer de steekproef voldoende representatief en aselect is, zijn er goede schatters te vinden van de populatieparameters. De genoemde voorbeelden zijn voorbeelden van zogenaamde *puntschatters*, omdat de betreffende populatieparameter door één getal, een punt op de getallenrechte, wordt geschat. Naast puntschattingen kennen we ook nog zogenaamde

intervalschattingen. Zo'n interval wordt bepaald door een ondergrens a en/of een bovengrens b waarvan met een zekere mate van *betrouwbaarheid* gezegd kan worden dat het de werkelijke waarde van de te schatten populatieparameter bevat. Deze betrouwbaarheid dient dan geïnterpreteerd te worden als de kans β dat men een interval berekend heeft dat de werkelijke waarde van de geschatte parameter bevat. Het bedoelde interval wordt dan ook meestal het 100β % *betrouwbaarheidsinterval* genoemd. Voor het berekenen van zo'n interval is een puntschatter nodig. In de volgende paragraaf zullen we dieper ingaan op het begrip betrouwbaarheidsinterval.

8.3 Intervalschattingen: betrouwbaarheidsintervallen

We zullen beginnen met een formele definitie van het begrip betrouwbaarheidsinterval.

Definitie
Onder het $100\beta\%$-betrouwbaarheidsinterval van de populatieparameter τ (Griekse letter *tau*) verstaan we een uit steekproefresultaten berekend interval waarvan gezegd kan worden dat er $100\beta\%$ kans bestaat dat het de werkelijke waarde van de populatieparameter τ bevat.

Merk op dat de betrouwbaarheid van een betrouwbaarheidsinterval de kans is dat het de werkelijke waarde van de parameter τ bevat en *niet* de kans dat de werkelijke waarde van de parameter τ in het interval ligt. Immers, de werkelijke waarde van τ is onbekend verondersteld, maar wel uniek bepaald, omdat hij betrekking heeft op de hele populatie; τ is dus niet 'stochastisch' (aan toeval onderhevig), zodat het begrip kans niet op τ zelf betrekking heeft. Het betrouwbaarheidsinterval daarentegen is wel stochastisch (zodat hieraan wel het begrip kans gekoppeld kan worden). Het is namelijk gebaseerd op de numerieke waarde van een schatting, die berekend wordt uit steekproefresultaten. En dus heeft het een stochastisch karakter (twee verschillende, doch even grote steekproeven, zullen vrijwel altijd twee verschillende puntschattingen opleveren).
Een betrouwbaarheidsinterval kan zowel tweezijdig zijn als links-eenzijdig of rechts-eenzijdig. Bij een tweezijdig $100\beta\%$-betrouwbaarheidsinterval berekent men zowel de ondergrens a als de bovengrens b van het interval. Bij een links-eenzijdig respectievelijk rechts-eenzijdig $100\beta\%$ betrouwbaarheidsinterval berekent men slechts de ondergrens a respectievelijk de bovengrens b. In alle drie de gevallen geldt dat er $100\beta\%$-kans bestaat dat het berekende interval de werkelijke waarde van de populatieparameter τ bevat. Hoewel er voor elke denkbare parameter van elke denkbare kansverdeling een betrouwbaarheidsinterval geconstrueerd kan worden, zullen we ons in dit boek beperken tot het berekenen van een betrouwbaarheidsinterval voor de parameters μ en σ van een normale verdeling en voor de parameter p van een binomiale verdeling die door een normale verdeling benaderd kan worden.

8.4 Intervalschattingen van het gemiddelde

Voor het maken van een intervalschatting voor het gemiddelde gaan we ervan uit dat de populatie normaal verdeeld is. We moeten een onderscheid maken tussen een populatie met *bekende* standaardafwijking en een populatie met een *onbekende* standaardafwijking.

8.4.1 De intervalschatting van het gemiddelde van een normale verdeling met een bekende standaardafwijking

Wanneer we voor het onbekende gemiddelde μ van een normale verdeling met bekende standaardafwijking σ een intervalschatting (een betrouwbaarheidsinterval) willen berekenen, dienen we eerst (zie de definitie) de beschikking te hebben over een puntschatting van μ. Daartoe nemen we uit de populatie die door de betreffende normale verdeling beschreven wordt, een steekproef van n stuks en berekenen hieruit de bedoelde puntschatting. Het ligt voor de hand hiervoor het gemiddelde $\bar{x}$ (numerieke waarde van $\bar{X}$) te kiezen.
Stel nu dat we in een steekproef van 25 stuks uit een normale verdeling met onbekend gemiddelde μ en bekende standaardafwijking $\sigma = 10$ voor de schatter $\bar{X}$ van μ de waarde $\bar{x} = 75$ hebben gevonden. We kunnen ons nu afvragen welke waarden van de onbekende μ nog aannemelijk te noemen zijn. Het zal duidelijk zijn dat μ in de buurt van 75 zal liggen, maar tussen welke onder- en bovengrens?
Laten we in eerste instantie eens onderzoeken of $\mu = 69$ kan zijn.

Volgens stelling 5 uit hoofdstuk 7 (gevolg van de centrale limietstelling) geldt dat de kansvariabele $\bar{X}$ normaal verdeeld is met gemiddelde $\mu_{\bar{X}} = \mu$ en standaardafwijking $\sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{n}}$.
We kunnen dus zeggen dat de getransformeerde variabele

$$U = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\dfrac{\sigma}{\sqrt{n}}} \tag{8.1}$$

standaardnormaal verdeeld is.

Wanneer $\mu = 69$ en $\sigma = 10$ kunnen we dus zeggen dat
$U = \dfrac{\bar{X} - 69}{\dfrac{10}{\sqrt{23}}} = \dfrac{\bar{X} - 69}{2}$ standaardnormaal verdeeld is.

Er geldt dan: $P(\bar{X} > 75) = P\left(U > \frac{75-69}{2}\right) = P(U > 3) = 0{,}0013$.
Dit betekent – zie de linkerhelft van figuur 8.1 – dat de kans om bij waarden van μ kleiner dan $\mu_1 = 69$ een waarde van $\bar{X}$ te vinden die groter is dan $\bar{x} = 75$, kleiner is dan 0,0013.
We zeggen dan dat bij $\bar{x} = 75$ de waarde van μ behoudens een onbetrouwbaarheid 0,0013 niet kleiner is dan $\mu_1 = 69$.

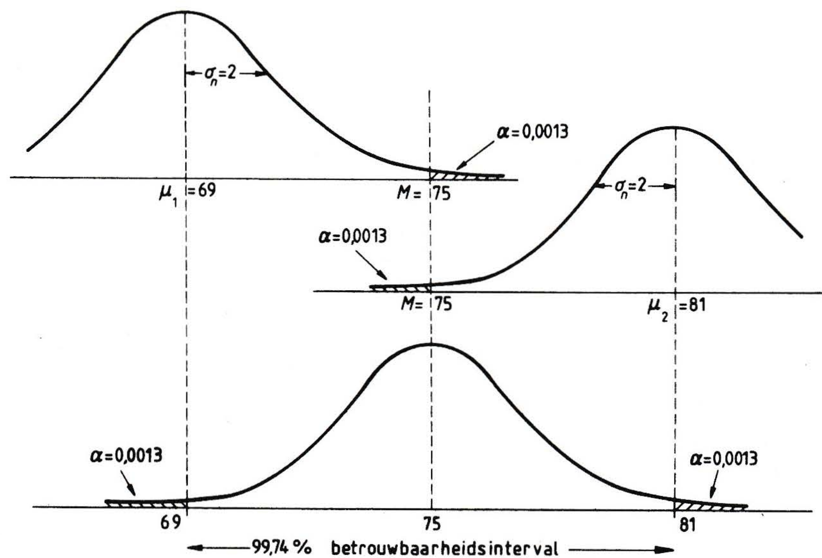


Fig. 8.1 Het 99,74%-betrouwbaarheidsinterval van μ bij $\sigma = 10$ met $n = 25$ en $\bar{x} = 75$

In tweede instantie onderzoeken we of $\mu = 81$ kan zijn. In dat geval geldt er – omdat $\mu_{\bar{X}} = 81$ en wederom $\sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{10}{\sqrt{25}} = 2$ – dat

$$U = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{X} - 81}{\frac{10}{\sqrt{23}}} = \frac{\bar{X} - 81}{2} \text{ standaardnormaal verdeeld is,}$$

zodat $P(\bar{X} < 75) = P\left(U < \frac{75-81}{2}\right) = P(U < -3) = P(U > 3) = 0{,}0013$.

Dit betekent – zie de rechterhelft van figuur 8.1 – dat de kans om bij waarden van μ groter dan $\mu_2 = 81$ een waarde van $\bar{X}$ te vinden die kleiner is dan $\bar{x} = 75$, kleiner is dan 0,0013. We zeggen dan dat bij $\bar{x} = 75$ de waarde van μ behoudens een onbetrouwbaarheid 0,0013 niet groter is dan $\mu_2 = 81$.

Combinatie van de beide bovenstaande uitspraken – zie wederom figuur 8.1 – leidt ertoe te zeggen dat bij $\bar{x} = 75$ de waarde van μ behoudens een onbetrouwbaarheid $2 \times 0{,}0013 = 0{,}0026$ niet kleiner is dan $\mu_1 = 69$ en niet groter is dan $\mu_2 = 81$.

We noemen de getallen $\mu_1 = 69$ en $\mu_2 = 81$ de grenzen van het 99,74%-betrouwbaarheidsinterval van μ. Wanneer we de positieve u-waarde die in de standaardnormale verdeling een rechteroverschrijdingskans α heeft aanduiden met $u(\alpha)$ geldt hier dus – met $\alpha = 0{,}0013$ en dus $u(\alpha) = 3$ – dat:

$$\mu_1 = 69 = 75 - 3 \cdot 2 = \bar{x} - u(\alpha) \cdot \frac{\sigma}{\sqrt{n}} \qquad \text{en}$$

$$\mu_2 = 81 = 75 + 3 \cdot 2 = \bar{x} + u(\alpha) \cdot \frac{\sigma}{\sqrt{n}}$$

We kunnen nu in het algemeen definiëren:

Definitie
Het $100\beta\%$-betrouwbaarheidsinterval van het gemiddelde μ van een normale verdeling met een bekende standaardafwijking σ wordt gegeven door de grenzen

$$\left[\bar{x} - u(\alpha) \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + u(\alpha) \cdot \frac{\sigma}{\sqrt{n}}\right] \tag{8.2}$$

waarin $\bar{x}$ het gemiddelde is van een steekproef van n stuks uit die normale verdeling en waarin $u(\alpha)$ het positieve getal is dat in de standaardnormale verdeling een rechteroverschrijdingskans $\alpha = \frac{1-\beta}{2}$ bezit.

Opmerking
De factor $\frac{\sigma}{\sqrt{n}}$ wordt in de literatuur ook wel *standaardfout* genoemd.

Opdracht

Bereken voor het behandelde voorbeeld het 95,44%-betrouwbaarheidsinterval van μ. Bepaal eveneens de betrouwbaarheid van het betrouwbaarheidsinterval [73, 77].

In tabel 8.1 hebben we op basis van tabel B1 voor enkele speciale gevallen de relatie tussen β, α en $u(\alpha)$ vastgelegd.

Tabel 8.1 De relatie tussen β, α en $u(\alpha)$ voor enkele waarden van β

$100\beta\%$	99	98	95	90	80	75	60	50
α	0,005	0,01	0,025	0,05	0,10	0,125	0,20	0,25
$u(\alpha)$	2,575	2,33	1,96	1,645	1,28	1,15	0,84	0,67

Merk op dat uit tabel 8.1 blijkt dat $u(\alpha)$ kleiner wordt – en dus het betrouwbaarheidsinterval smaller – naarmate β afneemt. Daarentegen wordt $u(\alpha)$ groter wordt – en dus het betrouwbaarheidsinterval breder – wanneer β toeneemt. De breedte van een betrouwbaarheidsinterval is bepalend voor de nauwkeurigheid ervan. Dit betekent dat de intervalschatting nauwkeuriger wordt naarmate men een kleinere betrouwbaarheid kiest en minder nauwkeurig wordt naarmate men een grotere betrouwbaarheid kiest.

Voorbeeld 1

Bij de fabricage van een bepaald soort glazen flessen wordt ervoor gezorgd dat de standaardafwijking van de normaal verdeelde breeksterkte van de flessen gelijk is aan 10 N(ewton). Van een steekproef van 16 flessen uit de dagproductie bleek de gemiddelde breeksterkte gelijk te zijn aan 110 N. Bereken een 95%-betrouwbaarheidsinterval voor de gemiddelde breeksterkte van de flessen in de dagproductie.

Oplossing

Met $\beta = 0{,}025$ vinden we in tabel 8.1: $u(\alpha) = 1{,}96$.

Voor het gevraagde 95%-betrouwbaarheidsinterval vinden we dan volgens formule (8.2):

$$\left[110 - 1{,}96 \times \frac{10}{\sqrt{16}};\ 110 + 1{,}96 \times \frac{10}{\sqrt{16}}\right] = [105{,}1;\ 114{,}9]$$

We zijn er dus voor 95% zeker van dat het populatiegemiddelde van de breekkrachten ligt tussen 105,1 N en 114,9 N. Deze conclusie kunnen we trekken omdat we weten dat bij herhaalde steekproeven (steeds van 16 breeksterkten) 95% van de intervallen, die op deze wijze geconstrueerd kunnen worden, het werkelijke populatiegemiddelde zal omvatten.

Opmerking

a. Ook wanneer de populatie (met onbekend gemiddelde μ en bekende standaardafwijking σ) niet normaal verdeeld is, kunnen we de gevolgde techniek gebruiken. Immers, ongeacht de soort verdeling waaruit de steekproef (mits voldoende groot) genomen wordt, kunnen we volgens stelling 5 uit hoofdstuk 7 voor het steekproefgemiddelde $\bar{X}$ de normale verdeling gebruiken.

b. In het voorafgaande is stilzwijgend verondersteld dat er sprake is van een steekproef met teruglegging of van een steekproef zonder teruglegging uit een 'oneindig grote' populatie. Wordt de steekproef echter zonder teruglegging genomen uit een eindige, ten opzichte van de steekproefomvang relatief kleine populatie met omvang N, dan moet in formule (8.2) de term $\frac{\sigma}{\sqrt{n}}$ volgens formule (7.13) vervangen worden door $\sqrt{\frac{N-n}{N-1}} \cdot \frac{\sigma}{\sqrt{n}}$

8.4.2 De intervalschatting van het gemiddelde van een normale verdeling met een onbekende standaardafwijking; de t-verdeling

In paragraaf 8.3.1 hebben we voor de constructie van het betrouwbaarheidsinterval van het gemiddelde μ van een normale verdeling met bekende standaardafwijking σ gebruikgemaakt van de standaardnormale verdeling van de kansvariabele $U = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{X} - \mu}{\sigma} \cdot \sqrt{n}$.

Wanneer de standaardafwijking σ niet bekend is, kunnen we het gegeven dat $U = \frac{\bar{X} - \mu}{\sigma} \cdot \sqrt{n}$ standaardnormaal verdeeld is, niet langer gebruiken.

Het ligt voor de hand in de uitdrukking $U = \dfrac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} = \dfrac{\bar{X}-\mu}{\sigma} \cdot \sqrt{n}$ de σ te vervangen door de geschatte standaardafwijking S van de steekproef (met waarde s). U gaat dan over in $\dfrac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} = \dfrac{\bar{X}-\mu}{S} \cdot \sqrt{n}$, maar van deze kansvariabele kan *niet* meer beweerd worden dat hij een standaardnormale verdeling bezit. De letter U is dan ook niet meer van toepassing. Men zou zich nu kunnen afvragen welke kansverdeling $T = \dfrac{\bar{X}-\mu}{s} \cdot \sqrt{n}$ dan wel bezit. Het is de statisticus William S. Gosset (1876-1937) die het antwoord op deze vraag gegeven heeft. Hij is de ontdekker van de kansverdeling van de kansvariabele

$$T = \frac{\bar{X}-\mu}{s} \cdot \sqrt{n} \tag{8.3}$$

als functie van het aantal *vrijheidsgraden* ν (Griekse letter, spreek uit: *nu*) van de standaardafwijking S . De *t-verdeling* is destijds onder de schuilnaam Student gepubliceerd.

Opmerking

We herinneren er nogmaals aan dat de naam van een kansvariabele met een hoofdletter geschreven wordt en de waarde met een kleine letter. Vandaar het onderscheid tussen T en t, het onderscheid tussen S en s en het onderscheid tussen U en u.

Het begrip vrijheidsgraad is in hoofdstuk 3 al even aan de orde geweest (bij de definitie van de standaardafwijking van een steekproef). Student's t-verdeling met ν vrijheidsgraden is dus de kansverdeling van de kansvariabele $T = \dfrac{\bar{X}-\mu}{S} \cdot \sqrt{n}$. Deze verdeling wordt volledig wordt bepaald door zijn parameter ν. De verdeling bezit een aantal eigenschappen die veel overeenkomst vertonen met die van de standaardnormale verdeling: voor elke waarde van ν is het gemiddelde van de verdeling gelijk aan 0, T kan waarden t aannemen tussen $-\infty$ en $+\infty$ en de verdeling is symmetrisch t.o.v. $t = 0$ (zie figuur 8.2).

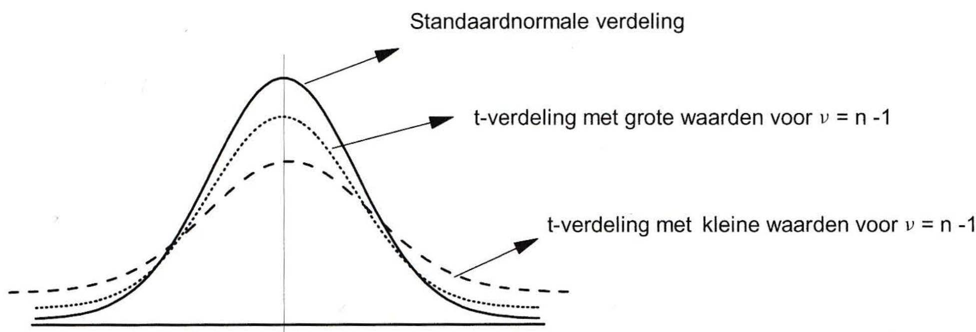


Fig. 8.2 Student's t-verdeling en de standaardnormale u-verdeling

Verder is de variantie van de verdeling voor elke waarde van $\nu > 2$ gelijk aan $\dfrac{\nu}{\nu - 2}$ waaruit volgt dat de variantie – en dus ook de standaardafwijking – tot 1 nadert naarmate ν groter wordt.
Redenerend zoals we in paragraaf 8.3.1 hebben gedaan rond figuur 8.1, kunnen we nu met behulp van de t-verdeling definiëren:

Definitie
Het $100\beta\%$-betrouwbaarheidsinterval van het gemiddelde μ van een normale verdeling met onbekende standaardafwijking σ wordt gegeven door de grenzen

$$\left[\bar{x} - t_\nu(\alpha) \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_\nu(\alpha) \cdot \frac{s}{\sqrt{n}}\right] \tag{8.4}$$

met $t_\nu(\alpha) > 0$
Hierin is $\bar{x}$ het gemiddelde en s de standaardafwijking van een steekproef van n stuks uit de betreffende normale verdeling. Verder is $t_\nu(\alpha)$ het positieve getal dat in de t-verdeling met $\nu = n - 1$ vrijheidsgraden een rechteroverschrijdingskans $\alpha = \dfrac{1 - \beta}{2}$ bezit.

Opdracht
Ga na dat uit de eigenschappen van de T-verdeling volgt dat voor elke waarde van ν de in formule (8.4) bedoelde factor $t_\nu(a)$ voor dezelfde waarde van α groter is dan de in formule (8.2) bedoelde factor $u(\alpha)$, maar dat dit in mindere mate het geval is naarmate ν groter is.

De tabel van de t-verdeling
Voor verschillende waarden van α en ν zijn waarden van de factor $t_\nu(\alpha)$ vastgelegd in tabel B5. Zo vinden we bijvoorbeeld in deze tabel bij $\alpha = 0{,}10$ en $\nu = 15$ de factor $t_{15}(0{,}10) = 1{,}341$, hetgeen betekent dat in de t-verdeling met 15 vrijheidsgraden de kans op een waarde van T groter dan 1,341 gelijk is aan 0,10. Merk op dat in verband met de symmetrie van de t-verdeling ten opzichte van $t = 0$ geldt dat $t_\nu(1 - \alpha) = -t_\nu(a)$. Zo is bijvoorbeeld $t_{25}(0{,}99) = -t_{25}(0{,}01) = -2{,}485$, hetgeen betekent dat in de t-verdeling met 25 vrijheidsgraden de kans op een waarde van T groter dan -2,485 gelijk is aan 0,99 en dus de kans op een waarde van T kleiner dan -2,485 gelijk is aan 0,01.

Opdracht
De t-verdeling met ν vrijheidsgraden gaat voor toenemende waarden van ν over in een standaardnormale verdeling. Ga met behulp van tabel B5 en tabel 8.1 na dat het gevolg hiervan is dat bij toenemende waarden van ν de waarde van de factor $t_\nu(\alpha)$ nadert tot die van de factor $u(\alpha)$.

Ga na dat uit het bovenstaande volgt dat het betrouwbaarheidsinterval volgens formule (8.4) bij toenemende waarde van de steekproefgrootte n nadert tot dat volgens formule (8.2).

Voorbeeld 2

Stel dat in het geval van voorbeeld 1 de standaardafwijking σ van de breeksterkte van de glazen flessen niet bekend is, maar dat de standaardafwijking van de breeksterkte van de 16 flessen in de genomen steekproef met gemiddelde $\bar{x} = 110$ N gelijk is aan $s = 10$ N. Bereken in dat geval een 95%-betrouwbaarheidsinterval voor de gemiddelde breeksterkte van de flessen in de dagproductie waaruit de betreffende steekproef afkomstig is.

Oplossing

Met $\beta = 0{,}95$ en dus $\alpha = \dfrac{1-\beta}{2}$ vinden we met $\nu = n - 1 = 15$ in tabel B5: $t_\nu(a) = t_{15}(0{,}025) = 2{,}131$. Voor het gevraagde 95%-betrouwbaarheidsinterval vinden we dan volgens formule (8.4):

$$\left[110 - 2{,}131 \times \frac{10}{\sqrt{16}};\ 110 + 2{,}131 \times \frac{10}{\sqrt{16}}\right] = [104{,}67;\ 115{,}33]$$

8.5 De intervalschatting van de variantie van een normale verdeling; de Chi-kwadraatverdeling

In de vorige twee paragrafen hebben we voor het maken van een intervalschatting (het berekenen van een betrouwbaarheidsinterval) van het gemiddelde μ van een normale verdeling met bekende en onbekende standaardafwijking σ gebruikgemaakt van de standaardnormale verdeling van de kansvariabele $U = \frac{\bar{X}-\mu}{\sigma} \cdot \sqrt{n}$ respectievelijk van de t-verdeling met $\nu = n - 1$ vrijheidsgraden van de kansvariabele $T = \frac{\bar{X}-\mu}{S} \cdot \sqrt{n}$.
Ook voor de berekening van een betrouwbaarheidsinterval van de variantie σ^2 van een normale verdeling beschikken we over een kansverdeling: *de χ^2-verdeling* (spreek uit: chi-kwadraatverdeling) met ν vrijheidsgraden.
De χ^2-verdeling met $\nu = n$ vrijheidsgraden wordt gedefinieerd als de kansverdeling van de kansvariabele $\chi^2 = \sum\limits_{i=1}^{n} U_i^2$ waarin U_i $(i = 1, 2, 3, ..., n)$ voor elke i standaardnormaal verdeeld is. Schrijven we voor $U_i = \dfrac{X_i - \mu}{\sigma}$ (met X_i $(i = 1, 2, 3, ..., n)$ voor elke i normaal verdeeld met gemiddelde μ en standaardafwijking σ), dan betekent dit dat de kansvariabele $\chi^2 = \sum\limits_{i=1}^{n} \left(\dfrac{X_i - \mu}{\sigma}\right)^2$ een χ^2-verdeling met $\nu = n$ vrijheidsgraden bezit. Wanneer we hierin μ vervangen door zijn schatter $\bar{X}$ (waardoor er een graad van vrijheid verloren gaat), dan betekent dit dat de kansvariabele

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{\sigma^2}$$

verdeeld is volgens een χ^2-verdeling met $\nu = n - 1$ vrijheidsgraden. Bedenken we dat met S^2 als schatter van σ^2 uit

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n - 1}$$

(formule voor de standaardafwijking van een steekproef, zie hoofdstuk 3) volgt dat

$$\sum_{i=1}^{n} (X_i - \bar{X})^2 = (n - 1)S^2$$

dan kan geconcludeerd worden dat de kansvariabele

$$\chi^2 = (n - 1)\frac{S^2}{\sigma^2} \tag{8.5}$$

een χ^2-verdeling met $\nu = n - 1$ vrijheidsgraden bezit.

De χ^2-verdeling lijkt op het oog weinig op een standaardnormale verdeling of een t-verdeling. De verdeling wordt (overigens evenals de t-verdeling) volledig bepaald door het aantal vrijheidsgraden ν van de schatter S^2 van σ^2.

χ^2 ligt voor elke waarde van ν tussen 0 en ∞ en de verdeling is binnen deze grenzen, zeker voor kleinere waarden van ν, sterk rechts-asymmetrisch. Voor toenemende waarden van ν neemt de asymmetrie van de verdeling af, voor voldoende grote ν wordt de verdeling zelfs min of meer symmetrisch.

Bewezen kan worden dat voor elke waarde van ν de χ^2-verdeling met ν vrijheidsgraden een gemiddelde (verwachtingswaarde) ν en een variantie 2ν heeft. Omdat de χ^2-verdeling voor toenemende waarden van ν steeds minder asymmetrisch wordt, mag voor voldoende grote waarden van ν (zeg $\nu > 30$) worden aangenomen dat de χ^2-verdeling met ν vrijheidsgraden redelijk goed benaderd kan worden door een normale verdeling met gemiddelde ν en standaardafwijking $\sqrt{2\nu}$.

In figuur 8.3 is voor enkele waarden van ν de kromme van de χ^2-verdeling weergegeven.

Om op basis van de berekende variantie s^2 van een steekproef van n stuks uit een normale verdeling met variantie σ^2 een $100\beta\%$-betrouwbaarheidsinterval voor σ^2 te construeren, redeneren we als volgt.

Wanneer χ_1^2 en χ_2^2 de waarden zijn die in de χ^2-verdeling met ν vrijheidsgraden een linker- respectievelijk rechteroverschrijdingskans $\frac{1 - \beta}{2}$ bezitten, dan geldt er, omdat $(n - 1)\frac{S^2}{\sigma^2}$ een χ^2-verdeling met ν vrijheidsgraden bezit:

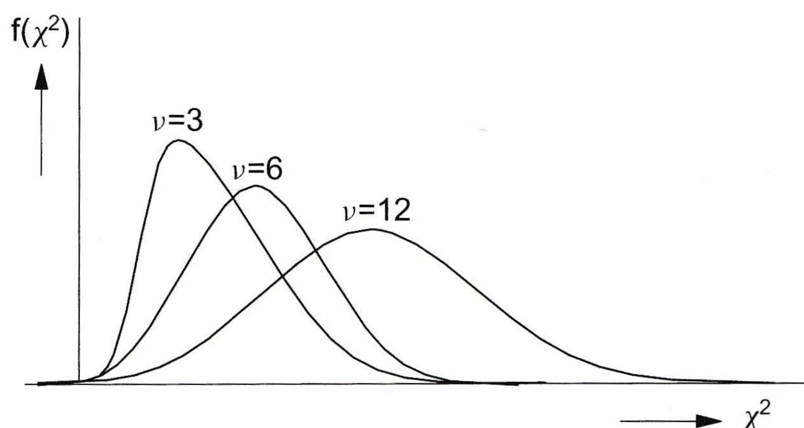


Fig. 8.3 Het verloop van de χ^2-verdeling voor enkele waarden van het aantal vrijheidsgraden ν

$$P\left(\chi_1^2 < (n-1)\frac{S^2}{\sigma^2} < \chi_2^2\right) = 1 - \beta$$

Wanneer we de drie leden van deze ongelijkheid delen door $(n-1)S^2$, ontstaat er een uitdrukking die – na herleiding – er als volgt uitziet:

$$P\left((n-1)\frac{S^2}{\chi_2^2} < \sigma^2 < (n-1)\frac{S^2}{\chi_1^2}\right) = 1 - \beta$$

Dit betekent dat er een kans $1-\beta$ bestaat dat het interval $\left[(n-1)\frac{s^2}{\chi_2^2}, (n-1)\frac{s^2}{\chi_1^2}\right]$ de werkelijke waarde σ^2 bevat (en dus is de kans β dat de werkelijke waarde van σ^2 buiten het betreffende interval ligt).
Met dit interval hebben we het $100\beta\%$-betrouwbaarheidsinterval voor de variantie σ^2 gevonden.
We kunnen nu in het algemeen definiëren:

Definitie
Het $100\beta\%$-betrouwbaarheidsinterval van de variantie σ^2 van een normale verdeling wordt gegeven door het interval:

$$\left[(n-1)\frac{s^2}{\chi_2^2}, (n-1)\frac{s^2}{\chi_1^2}\right] \tag{8.6}$$

Hierin is s^2 de (waarde van de) variantie van een steekproef van n stuks uit die normale verdeling.

Verder is $\chi_2^2 = \chi_\nu^2(\alpha)$ het getal dat in de χ^2-verdeling met $\nu = n - 1$ vrijheidsgraden een rechteroverschrijdingskans $\alpha = \dfrac{1-\beta}{2}$ bezit en $\chi_1^2 = \chi_\nu^2(1-\alpha)$ is het getal dat in de χ^2-verdeling met $\nu = n - 1$ vrijheidsgraden een rechteroverschrijdingskans bezit van $1 - \alpha = 1 - \frac{1-\beta}{2}$, dus een linkeroverschrijdingskans bezit van $\alpha = \frac{1-\beta}{2}$.

De tabel van de chi-kwadraat (χ^2)-verdeling

Voor verschillende waarden van α en ν zijn waarden van de factor $\chi_\nu^2(\alpha)$ vastgelegd in tabel B6.

In deze tabel vinden we bijvoorbeeld bij $\alpha = 0{,}10$ en $\nu = 15$ de factor $\chi_{15}^2(0{,}10) = 22{,}31$. Dit betekent dat in de χ^2-verdeling met 15 vrijheidsgraden de kans op een waarde van χ^2 groter dan 22,31 gelijk is aan 0,10. En bij $\alpha = 0{,}99$ en $\nu = 25$ is $\chi_{15}^2(0{,}99) = 11{,}52$. Dit betekent dat in de χ^2-verdeling met 25 vrijheidsgraden de kans op een waarde van χ^2 groter dan 11,52 gelijk is aan 0,99. De kans op een waarde van χ^2 kleiner dan 11,52 is gelijk aan 0,01.

Opdracht

Ga op grond van de definitie van de χ^2-verdeling na dat de kansvariabele U^2 een χ^2-verdeling met $\nu = 1$ vrijheidsgraad bezit en dat in verband hiermede voor elke waarde van α geldt dat $\chi_1^2(\alpha) = u^2(\frac{1}{2}\alpha)$.

Controleer dit voor enkele waarden van α met behulp van de tabellen B1 en B6.

Voorbeeld 3

Stel dat in het geval van voorbeeld 1 de standaardafwijking σ van de breeksterkte van de glazen flessen niet bekend is en dat de standaardafwijking van de 16 breeksterkten in de genomen steekproef met gemiddelde $\bar{x} = 110$ N gelijk is aan $s = 10$ N. Bereken in dat geval het 95%-betrouwbaarheidsinterval voor de variantie van de breeksterkten van de flessen in de dagproductie waaruit de betreffende steekproef afkomstig is.

Oplossing

Met $\beta = 0{,}95$ dus $\alpha = \dfrac{1-\beta}{2}$ vinden we met $\nu = n - 1 = 15$ in tabel B6:

$\chi_2^2 = \chi_\nu^2(\alpha) = \chi_{15}^2(0{,}025) = 27{,}49$ en

$\chi_1^2 = \chi_\nu^2(1-\alpha) = \chi_{15}^2(0{,}975) = 6{,}26$.

Voor het gevraagde 95%-betrouwbaarheidsinterval vinden we dan volgens formule (8.6):

$$\left[15 \times \frac{10^2}{27{,}49};\ 15 \times \frac{10^2}{6{,}26}\right] \text{ oftewel } [54{,}57;\ 238{,}62].$$

Wanneer – voor $\nu > 30$ – de χ^2-verdeling met ν vrijheidsgraden benaderd kan worden door een normale verdeling, kunnen de in formule (8.6) bedoelde factoren $\chi_2^2 = \chi_\nu^2(\alpha)$ en $\chi_1^2 = \chi_\nu^2(1-\alpha)$ vervangen worden door:

$\chi_2^2 = \mu + u(\alpha) \times \sigma = \nu + u(\alpha)\sqrt{2\nu}$ en $\chi_1^2 = \mu - u(\alpha) \times \sigma = \nu - u(\alpha)\sqrt{2\nu}$.

Voorbeeld 4

Stel dat in het geval van voorbeeld 1 de standaardafwijking van de breeksterkten in een steekproef van 51 flessen gelijk is aan s = 10 N. Bereken in dat geval het 95%-betrouwbaarheidsinterval voor de variantie van de breeksterkten van de flessen in de dagproductie waaruit de steekproef afkomstig is.

Oplossing

Omdat $\nu = n - 1 = 50$ groter is dan $\nu = 30$, geldt met $\beta = 0{,}95$
(dus $\alpha = \dfrac{1-\beta}{2} = 0{,}025$) dat:
$u(\alpha) = u(0{,}025) = 1{,}96$, dus:
$\chi_2^2 = 50 + 1{,}96 \cdot \sqrt{2 \times 50} = 69{,}6$ en
$\chi_1^2 = 50 - 1{,}96 \cdot \sqrt{2 \times 50} = 30{,}4$.
Voor het gevraagde 95%-betrouwbaarheidsinterval vinden we dan volgens formule (7.19):

$$\left[50 \times \frac{10^2}{69{,}6}; 50 \times \frac{10^2}{30{,}4}\right] = [71{,}84; 164{,}47]$$

Opmerking

De grenzen van het 100β%-betrouwbaarheidsinterval van de standaardafwijking σ van een normale verdeling zijn in principe *niet* de wortels van de grenzen van het 100β%-betrouwbaarheidsinterval van de variantie σ^2. Wanneer we toch op deze wijze een betrouwbaarheidsinterval van σ berekenen, dienen we ons te realiseren dat de betrouwbaarheid ervan geringer is dan die van het betrouwbaarheidsinterval van σ^2

8.6 De intervalschatting van een percentage

In hoofdstuk 5 hebben we het volgende gezien: Stel dat een populatie met N elementen een fractie p elementen met een bepaald kenmerk bevat. Dan is het aantal elementen K met dat kenmerk in een steekproef van n stuks uit die populatie binomiaal verdeeld met de parameters n en p. Dit geldt onder de voorwaarde dat de steekproef met teruglegging is. Wanneer de steekproef zonder teruglegging is, geldt dit ook, mits de populatie groot genoeg is ten opzichte van de steekproef (zeg $N > 10n$).
Wanneer bij $p \geq \frac{1}{2}$ geldt dat $n \geq 9\dfrac{p}{1-p}$ of bij $p \leq \frac{1}{2}$ geldt dat $n \geq 9\dfrac{1-p}{p}$ (zie paragraaf 6.5.1), kan de binomiale verdeling benaderd worden door een normale verdeling. Er geldt dan dat de kansvariabele K normaal verdeeld is met gemiddelde $\mu = np$ en standaardafwijking $\sigma = \sqrt{np(1-p)}$. Dit betekent dus dat de kansvariabele $\dfrac{K}{n}$ (= de fractie van het steekproefaantal met het bedoelde kenmerk) normaal verdeeld is met $\mu = p$ en standaardafwijking $\sigma = \dfrac{\sqrt{np(1-p)}}{n} = \sqrt{\dfrac{p(1-p)}{n}}$.

Wanneer $\dfrac{K}{n}$ normaal verdeeld is met gemiddelde $\mu = p$ en standaardafwijking $\sigma = \sqrt{\dfrac{p(1-p)}{n}}$, dan is de kansvariabele $U = \dfrac{\frac{K}{n} - p}{\sqrt{\frac{p(1-p)}{n}}}$ standaardnormaal verdeeld.

In dat geval geldt met een betrouwbaarheid van $100\beta\%$ dat $U = \dfrac{\frac{K}{n} - p}{\sqrt{\frac{p(1-p)}{n}}}$ ligt tussen $-u(\alpha)$ en $+u(\alpha)$:

$$-u(\alpha) < \frac{\frac{K}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} < +u(\alpha) \tag{8.7}$$

Formule (8.7) is na kwadratering ter herleiden tot de ongelijkheid

$$(n + u^2(\alpha))p^2 - (2K + u^2(\alpha))p + \frac{K^2}{n} < 0$$

Volgens de bekende theorie der kwadratische vergelijkingen is aan deze ongelijkheid voldaan voor die waarden van p, welke liggen tussen de beide oplossingen $p_1$en p_2 van de vergelijking die ontstaat als we het $<$-teken vervangen door een =-teken. Dus zijn de oplossingen van deze vergelijking de grenzen van het $100\beta\%$-betrouwbaarheidsinterval van p.

Definitie

Het $100\beta\%$-betrouwbaarheidsinterval van de parameter p van de binomiale verdeling van de kansvariabele K, is (wanneer deze verdeling benaderd kan worden door een normale verdeling) het interval $[p_1, p_2]$, waarbij p_1 en p_2 de oplossingen zijn van de vergelijking:

$$(n + u^2(\alpha))p^2 - (2k + u^2(\alpha))p + \frac{k^2}{n} = 0 \tag{8.8}$$

waarin $\alpha = \frac{1-\beta}{2}$ en k een schatting is van de kansvariabele K (op basis van een steekproef van n elementen).

Opmerking

Om te kunnen vaststellen of de binomiale verdeling van K benaderd kan worden door een normale verdeling, nemen we een steekproef van n elementen, bepalen vervolgens het aantal elementen k dat het bewuste kenmerk bezit. Dan kan vastgesteld worden dat de fractie van dat aantal gelijk is aan $\hat{p} = \dfrac{k}{n}$. Deze puntschatter van de populatiefractie

p dient te voldoen aan de voorwaarden, welke zojuist genoemd zijn: Als $\hat{p} > \frac{1}{2}$ moet gelden dat $n \geq 9\dfrac{\hat{p}}{1-\hat{p}}$ en als $\hat{p} < \frac{1}{2}$ moet gelden dat $n \geq 9\dfrac{1-\hat{p}}{\hat{p}}$.

Voorbeeld 5

Bereken een 95%-betrouwbaarheidsinterval voor het percentage ICT-deskundigen met een academische opleiding wanneer op een congres van academici 10 van de 125 aanwezigen ICT-deskundigen zijn.

Oplossing

Met $n = 125$ en $k = 10$ is $\hat{p} = \dfrac{10}{125} = 0{,}08$ en $1 - \hat{p} = \dfrac{115}{125} = 0{,}92$.

Er is dus voldaan aan de voorwaarde $n \geq 9\dfrac{1-\hat{p}}{\hat{p}} = 9 \cdot \dfrac{0{,}92}{0{,}08} = 103{,}5$.

Benadering van de binomiale verdeling met $n = 125$ en onbekende p is dus toegestaan.

Met $\alpha = \frac{0{,}05}{2} = 0{,}025$ dus $u(\alpha) = 1{,}96$ vinden we vergelijking (8.8). De coëfficiënten hiervan blijken te zijn:

$u^2(\alpha) + n = 128{,}84$

$u^2(\alpha) + 2k = 23{,}84$ en

$\dfrac{k^2}{n} = 0{,}8$.

De oplossingen van de kwadratische vergelijking $128{,}84p^2 - 23{,}84p + 0{,}8 = 0$ zijn gelijk aan

$p_1 = 0{,}0440$ en $p_2 = 0{,}1410$

Het 95%-betrouwbaarheidsinterval van het percentage ICT-deskundigen met een academische opleiding is dus $4{,}40 < p < 14{,}10$.

Opmerking

De geschetste methode is niet toegestaan wanneer K niet binomiaal (maar hypergeometrisch) verdeeld is of indien K wel binomiaal verdeeld is maar niet benaderd kan worden door een normale verdeling. In dat geval dienen de grenzen van het betrouwbaarheidsinterval van p berekend te worden met behulp van de hypergeometrische verdeling dan wel de binomiale verdeling. Het construeren van betrouwbaarheidsintervallen voor de parameters van dergelijke discrete kansverdelingen valt echter buiten het kader van dit boek.

8.7 Het bepalen van de steekproefgrootte voor het schatten van een gemiddelde

In de voorgaande paragrafen hebben we methoden besproken voor het berekenen van intervalschattingen (betrouwbaarheidsintervallen) van een gemiddelde, een variantie en een percentage (fractie). In alle gevallen die we daarbij beschouwden, werd de steekproefgrootte bekend verondersteld of was deze – in de voorbeelden – gegeven. In de praktijk

zullen we de steekproefgrootte vaak zelf moeten bepalen. De vraag rijst dan welke criteria daarbij gesteld moeten worden. Een belangrijk criterium is het kostenaspect. Het nemen van steekproeven brengt uiteraard kosten met zich mee: vaste kosten (de voorbereiding, de organisatie) en variabele kosten welke mede bepaald worden door het aantal steekproeven en de grootte van de steekproeven.

Een tweede belangrijk criterium is dat de steekproefgrootte voldoet aan de eisen voor het mogen toepassen van de beoogde berekeningsmethode voor het betrouwbaarheidsinterval. Zie de opmerkingen daarover in de voorgaande paragrafen. Ten slotte wordt de steekproefgrootte bepaald door de gewenste mate van betrouwbaarheid en de gewenste mate van *nauwkeurigheid* van de intervalschatting. Op dit laatste aspect zullen we in deze paragraaf ingaan. Daarbij beperken we ons (bij wijze van voorbeeld) tot het schatten van het gemiddelde van een normale verdeling met al of niet bekende standaardafwijking.

Wanneer we zeggen dat het $100\beta\%$-betrouwbaarheidsinterval van het gemiddelde μ van een normale verdeling met een bekende standaardafwijking σ een *nauwkeurigheid* ε moet bezitten, bedoelen we te zeggen dat het gemiddelde gelegen moet zijn op het interval

$$[\bar{x} - \varepsilon, \bar{x} + \varepsilon] \tag{8.9}$$

waarin $\bar{x}$ het gemiddelde is van een steekproef uit die normale verdeling.

In formule (8.1) vonden we het betrouwbaarheidsinterval voor het gemiddelde van een normale verdeling bij betrouwbaarheid β en $\alpha = \frac{1-\beta}{2}$:

$$\left[\bar{x} - u(\alpha) \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + u(\alpha) \cdot \frac{\sigma}{\sqrt{n}}\right]$$

Vergelijking met formule (8.9) geeft aan dat $\varepsilon = u(\alpha) \cdot \dfrac{\sigma}{\sqrt{n}}$ moet zijn. Anders gezegd: er moet voldaan zijn aan de formule

$$n = \frac{u^2(\alpha) \cdot \sigma^2}{\varepsilon^2} \tag{8.10}$$

waarin n de grootte is van de steekproef met het gemiddelde $\bar{x}$, terwijl $u(\alpha)$ met met $\alpha = \frac{1-\beta}{2}$ het positieve getal is dat in de standaardnormale verdeling een rechteroverschrijdingskans α heeft.

Uit formule (8.10) volgt dat bij een gegeven betrouwbaarheid de nauwkeurigheid van een intervalschatting des te groter is naarmate de steekproef groter is. En ook: hoe kleiner de steekproef, hoe kleiner de nauwkeurigheid.

Voor het gebruik van formule (8.10) is het noodzakelijk dat men weet hoe groot σ is. Voor het geval σ niet bekend is, kan deze ruwweg geschat worden door de spreidingsbreedte van een niet al te grote steekproef-vooraf door 6 te delen (immers de spreidingsbreedte van een normaalverdeelde variabele is ongeveer 6σ, want bijna alle waarnemingsuitkomsten

liggen tussen $\mu - 3\sigma$ en $\mu + 3\sigma$). Blijkt na invulling van de aldus verkregen schatting van σ in formule (8.10) de berekende steekproefomvang n groter te zijn dan die van de reeds genomen steekproef-vooraf, dan neemt men een aanvullende steekproef tot de berekende steekproefgrootte n bereikt is. Men beschouwt dan de beide steekproeven samen weer als steekproef-vooraf en start de procedure opnieuw. Dit doet men net zo lang (in de praktijk meestal 2 á 3 keer) totdat de omvang van alle reeds genomen steekproeven samen minstens gelijk is aan de berekende steekproefgrootte.

Voorbeeld 6

Een docent wil door middel van een steekproef het gemiddelde schatten van de scores (gehele getallen tussen 0 en 100) van een bepaald tentamen, waaraan 2000 studenten hebben deelgenomen. Hij wenst deze schatting bij een betrouwbaarheid van 95% te maken met een nauwkeurigheid van 2 scorepunten. Uit ervaring weet de docent dat tentamenscores bij goede benadering normaal verdeeld zijn. Hoeveel studenten moet de docent in zijn steekproef opnemen wanneer bekend is dat de laagst behaalde score 31 en de hoogst behaalde score 97 bedraagt?

Oplossing

Voor de standaardafwijking σ van de populatie kan, gezien de veronderstelling dat de scores bij benadering normaal verdeeld zijn, worden gekozen: $\sigma = \dfrac{97 - 31}{6} = 11$.
Met $\beta = 0{,}95$, dus $\alpha = 0{,}025$ en $u(\alpha) = 1{,}96$ vinden we met $\varepsilon = 2$ volgens formule (8.10):

$$n = \left(\frac{1{,}96 \times 11}{2}\right)^2 = 116{,}2.$$

Om de bedoelde schatting met de gewenste betrouwbaarheid en de gewenste nauwkeurigheid te kunnen maken, zal de docent van minstens 117 studenten de tentamenscore moeten vaststellen.

De geschetste methode is slechts geldig wanneer er sprake is van een steekproef met teruglegging of van een steekproef zonder teruglegging uit een voldoende grote populatie. Is hieraan niet voldaan, dus is er sprake van een steekproef zonder teruglegging uit een relatief kleine populatie, dan dient formule (8.10) lettend op formule (7.13) gecorrigeerd te worden.

Opgaven

1. **a.** Hoe groot is de kans om in de t-verdeling met 18 vrijheidsgraden een waarde t van T te vinden die ligt tussen $-2{,}101$ en $2{,}878$?

b. Hoe groot is de kans om in de t-verdeling met 60 vrijheidsgraden een waarde t van T te vinden die kleiner is dan $-2{,}000$ of groter is dan $2{,}660$?

c. Welke waarde t van T heeft in de t-verdeling met 10 vrijheidsgraden:
- een rechteroverschrijdingskans van 5%?
- een linkeroverschrijdingskans van 2,5%?
- een rechteroverschrijdingskans van 90%?
- een linkeroverschrijdingskans van 99%?

2. **a.** Hoe groot is de kans om in de χ^2-verdeling met 10 vrijheidsgraden een waarde van χ^2 te vinden die groter is dan 18,31?

b. Hoe groot is de kans om in de χ^2-verdeling met 24 vrijheidsgraden een waarde van χ^2 te vinden die kleiner is dan 13,85?

c. Hoe groot is de kans dat de waarde van χ^2 in de χ^2-verdeling met 72 vrijheidsgraden kleiner is dan 60 of groter is dan 80?

d. Welke waarde van χ^2 heeft in de χ^2-verdeling met 18 vrijheidsgraden:
- een rechteroverschrijdingskans van 5%?
- een linkeroverschrijdingskans van 1%?

e. Welke waarde van χ^2 heeft in de χ^2-verdeling met 50 vrijheidsgraden:
- een rechteroverschrijdingskans van 2,5%?
- een linkeroverschrijdingskans van 10%?

3. Het aantal lucifers per doosje is bij benadering normaal verdeeld met een standaardafwijking van 10 stuks. In een steekproef van 10 doosjes trof men achtereenvolgens 112, 101, 105, 119, 95, 104, 98, 110, 100 en 97 lucifers aan.

a. Bereken een 90%-betrouwbaarheidsinterval voor het gemiddelde aantal lucifers per doosje.

b. Neem aan dat σ niet bekend is en beantwoord nogmaals vraag a.

4. Het hoofd van de kwaliteitsdienst van een fabriek waar condensatoren vervaardigd worden, wil op basis van een steekproef van 250 condensatoren met een gemiddelde capaciteit van 90 Farad, met een betrouwbaarheid van 95% een intervalschatting maken van de gemiddelde capaciteit van de condensatoren in de partij waaruit die steekproef afkomstig is. Neem aan dat de standaardafwijking van de 6000 condensatoren in de bedoelde partij bekend is: deze bedraagt 5 Farad.
Bereken het gewenste 95%-betrouwbaarheidsinterval.

5. Om de sterkte van een bepaald type fietsband te bepalen, werden 9 banden zo ver opgepompt, dat ze kapot sprongen. De druk waarbij dit gebeurde bleek achtereenvolgens te zijn:

8,2	9,6	11,0	8,9	9,1	10,4	10,2	8,6	9,4

Wanneer aangenomen mag worden dat deze druk, waarbij opgepompte fietsbanden van dit type kapot springen ('maximumdruk' genoemd), normaal verdeeld is, bereken dan een 99%-betrouwbaarheidsinterval voor de gemiddelde maximumdruk.

6. Wanneer bij het testen van een bepaalde eigenschap van 10 exemplaren van een bepaald product de standaardafwijking van de 10 gevonden meetwaarden $s = 2$ bedraagt, bereken dan een 90%-betrouwbaarheidsinterval voor de variantie van de normaal verdeelde waarden van de betreffende producteigenschap.

7. Een audiometrist heeft bij 49 aselect gekozen medewerkers van bedrijven met veel omgevingslawaai de reactietijd op een bepaald auditief signaal gemeten. Voor het gemiddelde van de 49 reactietijden vond hij $\bar{x} = 0{,}7$ seconden.

- **a.** Wanneer uit vroeger onderzoek bekend is dat de bedoelde reactietijd normaal verdeeld is met standaardafwijking $\sigma = 0{,}2$ seconden, bereken dan een 95%-betrouwbaarheidsinterval voor de gemiddelde reactietijd van alle medewerkers uit de bedoelde bedrijven.
- **b.** Welk minimum aantal medewerkers moet de audiometrist bij zijn onderzoek betrekken wanneer hij de gemiddelde reactietijd wil schatten met een betrouwbaarheid van 95% en een nauwkeurigheid van 0,02 seconden?
- **c.** Wanneer uit vroeger onderzoek wel de normaliteit maar niet de standaardafwijking van de reactietijden bekend is, beantwoord dan nogmaals vraag a wanneer van de daar bedoelde 49 reactietijden de standaardafwijking $s = 0{,}2$ seconden bedraagt.

8. Bij een stembusenquête, enige tijd voor de verkiezing van de Tweede Kamer, zeiden 300 van de 900 ondervraagde kiesgerechtigden dat ze zouden gaan stemmen op partij A. Bereken een 95%-betrouwbaarheidsinterval voor het percentage kiesgerechtigden, dat van plan is op partij A te gaan stemmen.

9. Ten aanzien van een bepaald productieproces wordt de eis gesteld dat dit bij een juiste afstelling van het gemiddelde een standaardafwijking $\sigma = 5$ heeft.

- **a.** Wanneer een steekproef van 16 producten een standaardafwijking $s = 6{,}5$ bezit, bereken dan een 90%-betrouwbaarheidsinterval voor de variantie van het productieproces.
- **b.** Is er, behoudens een onbetrouwbaarheid van 10%, reden om aan te nemen dat het productieproces niet de juiste standaardafwijking heeft?

10. Bij een MMO-onderzoek (Multi Moment Opnamen) van een machinepark vond men na 3600 waarnemingen de volgende cijfers: in bedrijf 68%, stilstand 16%, reparatie 6% en onderhoud 10%. Bereken voor elk van deze 4 posten een 99%-betrouwbaarheidsinterval.

9 Het toetsen van hypothesen

9.1 Inleiding

In het dagelijks leven worden veel beweringen en veronderstellingen gedaan. Van sommige beweringen kan vrij eenvoudig worden nagaan of deze al of niet juist zijn.
Als iemand bijvoorbeeld zegt dat een bepaald pad langer is dan 100 m, kan door nameting eenvoudig worden nagaan of de bewering juist is of niet.
Als iemand zegt dat route A naar zijn werk korter is dan route B, kunnen we door de beide routes te rijden nagaan of deze bewering juist is. Moeilijker wordt het als de bewering als volgt luidt: 'Gemiddeld ben ik via route A sneller op mijn werk dan via route B'. Door allerlei oorzaken zijn er dagelijks variaties in de tijd die over beide routes gedaan wordt. Er is dus spreiding in de tijd en dit geeft problemen bij het doen van uitspraken over het al of niet waar zijn van de bewering.
Om toch iets te kunnen zeggen over dit soort van beweringen, komen we op het terrein van de statistische toetsen.
We geven twee inleidende voorbeelden.

Voorbeeld 1

Met een proef beoogde iemand te onderzoeken of het behandelen van een bepaalde rubbersoort met een chloorhoudende stof de slijtweerstand van rubber vergroot. De onderzoeker nam uit een partij aselect 10 proefstukjes van het rubber en verdeelde elk proefstukje in tweeën. De ene helft werd behandeld met de chloorhoudende stof en de andere helft werd onbehandeld gelaten. De keuze van de te behandelen helft werd overgelaten aan het lot (bijvoorbeeld door het werpen van een munt). De slijtweerstand van de 10 monsterparen (behandeld en onbehandeld) werd op een apparaat gemeten. De 10 verschillen in slijtweerstand, behandeld minus onbehandeld, zijn in tabel 9.1 gegeven.
De waarde van het gemiddelde verschil $\bar{v} = \dfrac{12{,}7}{10} = 1{,}27$ is positief. Dit suggereert dat de behandeling met de chloorhoudende stof gunstig is. De spreiding tussen de individu-

Tabel 9.1 Verschil in slijtweerstand van proefstukjes rubber

proefstuk nr	verschil V behandeld - onbehandeld
1	2,6
2	3,1
3	-0,2
4	1,7
5	0,6
6	1,2
7	2,2
8	1,1
9	-0,2
10	0,6
$\sum_{i=1}^{10}$	12,7

ele resultaten is echter vrij groot. Om nu een betrouwbare conclusie te kunnen trekken, doet men een zogenaamde *significantietoets*. Later zullen we zien hoe deze verloopt.

Voorbeeld 2

Een bepaalde kwaal kan behandeld worden met medicijn A. Uit ervaring is bekend dat in 50% van de gevallen de klachten van de kwaal na drie dagen zijn verholpen. Er is nu een nieuw medicijn B ontwikkeld, waarbij uit voorstudies lijkt alsof dit medicijn B effectiever is. Om een beslissing te nemen over de effectiviteit van medicijn B, wordt er een proef gedaan waarbij aan 100 patiënten medicijn B wordt toegediend. Het aantal patiënten, waarbij de klachten van de kwaal na drie dagen zijn verdwenen, wordt geteld. Op basis van dit aantal zal een beslissing worden genomen over de effectiviteit van medicijn B ten opzichte van medicijn A. Maar bij welk aantal 'herstelden' kan aangenomen worden dat medicijn B werkzamer is dan medicijn A? Het zal duidelijk zijn dat, indien bij 50 van de 100 patiënten de klachten verdwijnen, we niet besluiten dat B beter is dan A. De vraag is nu: 'hoeveel meer dan 50 zijn er nodig'? Is 51 al voldoende, of 55, 60, 75? Dus hoeveel patiënten moeten binnen drie dagen hersteld zijn, voordat geconcludeerd kan worden dat medicijn B beter is dan medicijn A? Ook hier zal een statistische toets uitkomst moeten bieden. We komen hier later op terug.

9.2 Theorie van het toetsen

Bij de beide gegeven voorbeelden willen we een antwoord hebben op de vraag:

'Kan het gevonden gemiddelde verschil worden toegeschreven aan toevallige oorzaken, of is het gevonden verschil groter dan op grond van het toeval mag worden verwacht?'
Kortgezegd willen we weten of we te maken hebben met een *toevallige afwijking*, of met een *systematische afwijking*.
In dit laatste geval spreekt men van een *significante* of aantoonbare afwijking.
In voorbeeld 1 vragen we ons af of de waarde van het gevonden gemiddelde verschil ($\bar{v} = 1{,}27$) significant afwijkt van $\mu_V = 0$. Als dit inderdaad zo is, is er geen verschil tussen behandeld en onbehandeld rubber ten aanzien van de slijtweerstand.
In voorbeeld 2 zullen er door toevallige oorzaken schommelingen in het aantal herstelden zijn in verschillende steekproeven. We willen een criterium aanleggen waarop we besluiten dat medicijn B wel of niet beter is dan medicijn A.
In de twee gegeven voorbeelden kunnen we 4 situaties onderscheiden waarin we terecht kunnen komen bij het nemen van een beslissing.

1. We beslissen dat chloorbehandeling een verbetering geeft ten aanzien van de slijtweerstand, terwijl dit in werkelijkheid niet het geval is (voorbeeld 1)
We beslissen dat het medicijn B effectiever is dan medicijn A, maar in werkelijkheid is dit niet het geval (voorbeeld 2) .
2. Chloorbehandeling geeft geen verbetering ten aanzien van de slijtweerstand en we beslissen dat ook (voorbeeld 1).
Medicijn B is niet effectiever dan medicijn A en we beslissen dat ook (voorbeeld 2).
3. We beslissen dat chloorbehandeling geen verbetering geeft ten aanzien van de slijtweerstand, terwijl dit wel het geval is (voorbeeld 1).
We beslissen dat medicijn B niet effectiever is dan medicijn A, terwijl dit wel het geval is (voorbeeld 2).
4. Chloorbehandeling geeft een verbetering ten aanzien van de slijtweerstand en we beslissen dat ook (voorbeeld 1).
Medicijn B is effectiever dan medicijn A en we beslissen dat ook (voorbeeld 2).

In de situaties 1 en 3 hebben we (ongewild) een beslissing genomen die niet met de werkelijkheid overeenkomt. Bij de situaties 2 en 4 hebben we een beslissing genomen die wel met de werkelijkheid overeenstemt. We moeten er nu voor zorgen dat we zo weinig mogelijk in situaties 1 en 3 komen te verkeren.
In situatie 1 spreken we in de statistiek van een *fout van de eerste soort* en in situatie 3 spreken we van een *fout van de tweede soort.*
Bij het toetsen beginnen we te veronderstellen dat er niets aan de hand is, dus dat er geen verschillen zijn tussen de bestaande situatie en de 'nieuwe' situatie. Deze veronderstelling vooraf noemt men de *nulhypothese*, aangegeven door het symbool H_0.
Naast de nulhypothese kennen we de *alternatieve hypothese*, aangeduid door het symbool H_1. In de alternatieve hypothese wordt vaak het tegengestelde van de nulhypothese aangegeven, vaak datgene wat juist aangetoond moet worden.
We nemen nu verder voorbeeld 1 als uitgang voor de bespreking van de toetsingstheorie.

Voor dit voorbeeld kunnen we de volgende hypothesen opstellen (de nulhypothese, respectievelijk de alternatieve hypothese):

$$H_0 \ : \ \mu_V = 0$$
$$H_1 \ : \ \mu_V \neq 0$$

Gezien het feit dat een steekproefgemiddelde $\overline{v}$, door toevalsvariaties, niet precies overeenkomt met het veronderstelde populatiegemiddelde $\mu_V (= 0)$, kan bij een 'afwijking' van $\overline{v}$ ten opzichte van μ_V niet zonder meer geconcludeerd worden dat er in werkelijkheid ook verschillen zijn. De waarde $\overline{v} = 1{,}27$ kan aanleiding geven tot de bewering dat de procesverandering (behandelen met een chloorhoudende stof) invloed heeft op de slijtweerstand. Met een dergelijke bewering lopen we een zeker risico. Het risico is hier dat er in werkelijkheid geen verschil is en dat slechts door *toeval* de afwijkende waarde van $\overline{v} = 1{,}27$ ten opzichte van $\mu_V = 0$ tot stand is gekomen.
Wil men bij een bewering het risico van een foute bewering aangeven, dan is het wenselijk daarvoor een getalswaarde in te voeren. Het is gebruikelijk dit als volgt te doen.
Het risico van de bewering *'er is een reële verandering opgetreden'* wordt vertaald in de kans dat een waarnemingsresultaat als het gevondene louter toevallig is ontstaan. Deze kans wordt de *overschrijdingskans* genoemd. Om de overschrijdingskans van de meetwaarde van de betreffende variabele te kunnen berekenen, moet de kansverdeling van de variabele bekend zijn. Daarbij wordt aangenomen dat de genoemde vooronderstelling (nulhypothese) waar is dat er geen verandering heeft plaatsgehad. Deze aanname geeft ons de parameters van de kansverdeling. De kansvariabele waarop de bewering gebaseerd wordt heet de *toetsingsvariabele*.
Nu keren we terug naar voorbeeld 1. Gezien de aard van de toets dient de toetsingsvariabele in dit geval een maat voor de ligging van het kenmerk 'verschil in slijtweerstand' te zijn. We kiezen hiervoor het gemiddelde verschil in behandelingsresultaat (naam $\overline{V}$ met waarde $\bar{v}$).
Naarmate de gevonden overschrijdingskans kleiner is, hebben we minder vertrouwen in H_0 en zijn we meer geneigd deze te verwerpen. Men verwerpt H_0 dan, als de overschrijdingskans onder of op een van tevoren vastgestelde grenswaarde komt. Die gekozen grenswaarde wordt de *onbetrouwbaarheidsdrempel* (of kortweg *onbetrouwbaarheid*) genoemd en wordt aangeduid met het symbool α.
Het risico van de bewering *'er is een reële verandering opgetreden'*, terwijl er in werkelijkheid geen verandering heeft plaatsgevonden, is dus de kans om H_0 ten onrechte te verwerpen. Dit is de maximale kans op het maken van een *fout van de eerste soort*.
Als onbetrouwbaarheidsdrempel kiest men meestal de waarde 0,05 ($\alpha = 0{,}05$). Men heeft dan een kans van 5% op een fout van de eerste soort. Anders gezegd: als de nulhypothese juist is en in een groot aantal gevallen getoetst wordt, zal in gemiddeld 1 op de 20 gevallen een waarde voor de toetsingsvariabele worden gevonden, waarbij H_0 ten onrechte

verworpen wordt (en dus een fout van de eerste soort gemaakt wordt).
Als H_0 wordt verworpen, doen we uitspraken als:

- er is een *reële* verhoging (of verlaging) geconstateerd;
- het gevonden gemiddelde ligt *systematisch* hoger (of lager);
- de gevonden verhoging (of verlaging) is *significant*.

Opmerking

a. Het niveau van de overschrijdingskans wordt vaak aangegeven met één of meer sterretjes. Bijvoorbeeld:
$0{,}01 < P < 0{,}05$ *
$0{,}001 < P < 0{,}01$ **
$P < 0{,}001$ ***

b. De formulering 'onder aanname van de nulhypothese H_0' wordt meestal verkort tot 'onder H_0'.

9.2.1 Fout van de eerste soort versus fout van de tweede soort

Uit het voorgaande blijkt dat als de bij de waarde van toetsingsvariabele gevonden overschrijdingskans groter is dan de gekozen onbetrouwbaarheidsdrempel α, H_0 niet wordt verworpen. Dit betekent echter *niet* dat H_0 dan ook juist moet zijn. Het *niet verwerpen* betekent slechts dat er geen reden is H_0 onjuist te achten en dat men de nulhypothese daarom aanvaardt (vergelijkbaar met een verdachte die vrijgesproken wordt wegens gebrek aan bewijs). Het is dan ook gebruikelijk om bij niet-verwerpen de conclusie voorzichtig te formuleren, bijvoorbeeld: *een reële verandering kon niet worden aangetoond op grond van het beschikbare waarnemingsmateriaal.*

In figuur 9.1 is de linker kansverdeling de (normale) verdeling van (de waarden van) de toetsingsvariabele, onder aanname dat de nulhypothese $H_0 : \mu = \mu_0$ juist is. Wanneer bij de (in een steekproef) gevonden waarde van de toetsingsvariabele een rechteroverschrijdingskans hoort die kleiner is dan de vooraf gestelde α, wordt de nulhypothese verworpen. Dit is het geval wanneer de gevonden waarde rechts van de waarde c op de horizontale as in figuur 9.1 ligt. In het geval dat de waarde van de toetsingsvariabele daadwerkelijk groter is dan c, wordt H_0 verworpen, terwijl H_0 blijkbaar toch juist kan zijn. Het risico op het nemen van deze foute beslissing is dus maximaal gelijk aan α.
Uit het voorgaande blijkt dat ook aan het niet-verwerpen van H_0 een risico is verbonden. Indien het werkelijke gemiddelde anders ligt dan in H_0 verondersteld is, is het toch mogelijk dat H_0 niet verworpen wordt. Het ten onrechte niet-verwerpen van H_0 wordt aangeduid als een *fout van de tweede soort* β.

De samenhang tussen α en β

De samenhang tussen α en β wordt in figuur 9.1 weergegeven. Stel eens dat het werkelijke gemiddelde veel groter is dan μ_0. De nulhypothese ($\mu = \mu_0$) is dan niet waar, dus is de

alternatieve hypothese $H_1 : \mu \geq \mu_0$ waar. De toetsingsvariabele heeft dan bijvoorbeeld de rechterverdeling ($\mu = \mu_1$) in figuur 9.1. Wanneer de met een steekproef gevonden waarde van de toetsingsvariabele in dat geval links ligt van $x = c$, zouden we ten onrechte de conclusie trekken dat de nulhypothese waar is. De linkeroverschrijdingskans van de toetsingsvariabele is dan kleiner dan β, de fout van de tweede soort. Uit figuur 9.1 is op te maken dat de kans op het maken van een fout van de tweede soort kleiner zal zijn naarmate het 'werkelijke' gemiddelde verder rechts van μ_0 (= gemiddelde onder H_0) ligt. Immers bij gelijkblijvende α zal de rechtse normale verdeling verder naar rechts opschuiven waardoor het hieronder, links van $x = c$ gelegen oppervlak β kleiner zal worden.

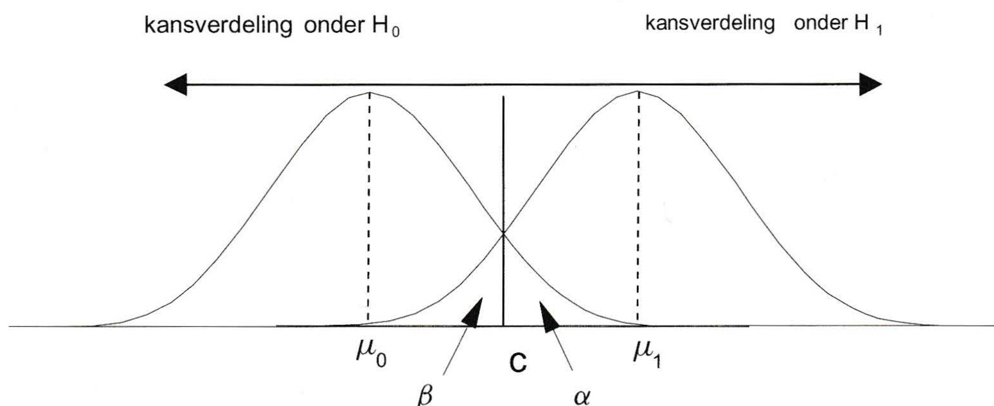


Fig. 9.1 Samenhang tussen fout van de eerste soort (α) en van de tweede soort (β)

Verder is β afhankelijk van de keuze van de onbetrouwbaarheidsdrempel α. Als men bijvoorbeeld α van 0,05 verkleint naar 0,01 (waardoor c in figuur 9.1 opschuift naar rechts), neemt de kans op het niet ontdekken van een afwijkend gemiddelde ten opzichte van μ_0 toe. Wanneer (nog steeds in figuur 9.1) het werkelijke gemiddelde gelijk is aan dezelfde μ_1, zal de linkeroverschrijdingskans β groter worden. Er wordt dus eerder een fout van de tweede soort gemaakt.

We kunnen (bij vaste steekproefgrootte) de kans op een fout van de ene soort verkleinen, maar daarbij wordt echter de kans op een fout van de tweede soort vergroot. Ten slotte hangt de fout van de tweede soort nog af van de steekproefgrootte n. Bij vaste α en bij toenemende n zal de kans op een fout van de tweede soort (β) kleiner worden. Bij toenemende n zal de toevallige afwijking van de toetsingsvariabele ten opzichte van het gemiddelde kleiner worden. We zullen hier later dieper op ingaan.

Bij de keuze van de onbetrouwbaarheidsdrempel en de grootte van n dient men de ernst van de verschillende mogelijke risico's tegen elkaar af te wegen. Uiteraard spelen kostenoverwegingen hierbij een rol. In de volgende paragraaf zullen we de toetsingsprocedure nader uiteenzetten.

9.2.2 Algemene gang van zaken bij het toetsen van hypothesen (toetsingsprocedure)

Omdat bij alle toetsen steeds dezelfde stappen moeten worden doorlopen, gaan we de gang van zaken bij het toetsen punt voor punt in een zogenaamde *toetsingsprocedure* samenvatten.

1. *Probleemstelling*
   De gestelde vraag wordt geanalyseerd. Wat wil de onderzoeker precies weten en wat wil hij aantonen? Uit de probleemstelling moet blijken wat getoetst moet worden en in welke afwijkingen hij geïnteresseerd is.

2. *Opstellen van de nulhypothese*
   Deze is veelal een ontkenning van wat de onderzoeker wil aantonen. Vermoedt hij bijvoorbeeld dat een gemiddelde groter is dan 25, dan luidt de nulhypothese: het gemiddelde is gelijk (of $\leq$) aan 25. Indien H_0 is geformuleerd, ligt ook H_1 vast. In H_1 wordt dan verwoord wat de onderzoeker wil aantonen. In de formulering van de nulhypothese komt altijd het $=$-teken voor.

3. *Keuze van de onbetrouwbaarheidsdrempel* α
   Doorgaans neemt men $\alpha = 0{,}05$, tenzij er een speciale reden is om een hogere of een lagere waarde te kiezen.

4. *Keuze van de toets*
   Het is gebruikelijk de toetsen te onderscheiden naar aard van de toetsingsvariabele (of toetsingsgrootheid). Soms bestaan er verschillende toetsen voor een bepaalde hypothese H_0. Een belangrijk criterium bij de keuze daartussen wordt gevormd door de kans op een fout van de tweede soort bij de mogelijke alternatieve hypothesen. Deze moet klein zijn bij die alternatieve hypothese die men het belangrijkst acht. Is dat bij een bepaalde toetsingsvariabele het geval, dan noemt men het *onderscheidingsvermogen* ($= 1 - \beta$) van de betreffende toets tegen die hypothesen groot. Verder wordt de keuze van de toets bepaald door:
   - de verdeling van de uitkomsten (normale of een andere verdeling);
   - wat getoetst moet worden (gemiddelden, spreidingen, enzovoorts).

5. *Uitvoering van de toets*
   De waarde van de toetsingsvariabele wordt berekend uit de waarnemingsuitkomsten.

6. *Bepaling overschrijdingskans*
   De overschrijdingskans, onder de nulhypothese, van de berekende toetsingsvariabele wordt bepaald. Hiervoor wordt de kansverdeling van de toetsingsvariabele gebruikt (meestal met behulp van tabellen of met een programma zoals EXCEL).

7. *Statistische conclusie*

   De gevonden overschrijdingskans (P) van de toetsingsvariabele (zie punt 6) wordt vergeleken met de gekozen onbetrouwbaarheidsdrempel α (zie punt 3). Indien $P < \alpha$ wordt H_0 verworpen. Als $P \geq \alpha$, wordt H_0 niet verworpen. De conclusie kan stelliger zijn naarmate de overschrijdingskans verder van de onbetrouwbaarheidsdrempel α af ligt.

8. *Technische conclusie*

   We moeten nu nog vertalen wat de statistische conclusie betekent. Dit is het antwoord op de in punt 1 gestelde vraag.

Aanvullende opmerkingen ten aanzien van het toetsen

Bij alle toetsen gaan we, evenals in de rechtspraak, ervan uit dat 'de verdachte' onschuldig is, met andere woorden: de nulhypothese geeft die situatie weer waarbij niets afwijkend aan de hand is. Vandaar dat in de nulhypothese altijd het =-teken voorkomt ($\geq$ of $\leq$ bevat ook een =-teken).

Voorbeeld 3

Als vervolg op voorbeeld 1 beginnen we met de veronderstelling, dat chloorbehandeling van rubber geen invloed heeft op de slijtweerstand (dat wil zeggen er is geen verschil tussen voor en na de behandeling: het gemiddeld verschil is 0). We kunnen dan de volgende hypothesen toetsen:

$H_0 : \mu_V = 0$ tegen $H_1 : \mu_V \neq 0$ (tweezijdig)
óf $H_1 : \mu_V < 0$ (linkseenzijdig)
óf $H_1 : \mu_V > 0$ (rechtseenzijdig)

Voorbeeld 4

Als vervolg op voorbeeld 2 veronderstellen we dat medicijn B minder dan of even effectief is als medicijn A. De hypothesen luiden dan:

H_0: $p_B \leq p_A$ tegen $H_1 : p_B > p_A$. Hierin is p_A de fractie van het aantal zieke personen dat na drie dagen met gebruik van medicijn A genezen is, idem voor p_B.

Voorbeeld 5

Wil men een machineafstelling controleren ten aanzien van een normwaarde, dan luiden de hypothesen:

H_0: $\mu = \mu_0$ (μ_0 = normwaarde) tegen H_1: $\mu \neq \mu_0$ òf H_1: $\mu < \mu_0$ òf H_1: $\mu > \mu_0$

Voorbeeld 6

Willen we de spreidingen van twee machines vergelijken door middel van varianties, dan veronderstellen we bij de nulhypothese dat de beide machines dezelfde spreiding hebben, dus:

H_0: $\sigma_A^2 = \sigma_B^2$ tegen $H_1 : \sigma_A^2 \neq \sigma_B^2$ òf H_1: $\sigma_A^2 < \sigma_B^2$ òf $H_1 : \sigma_A^2 > \sigma_B^2$

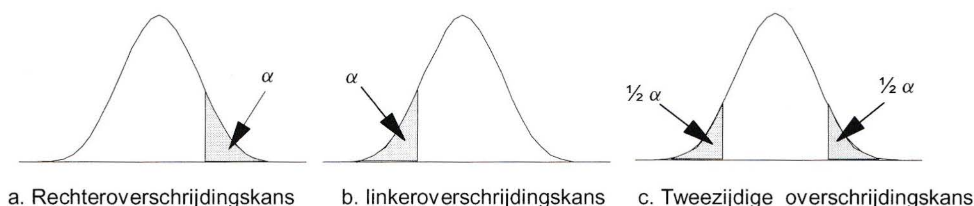
Uit bovenstaande voorbeelden blijkt dat er verschillende alternatieve hypothesen mogelijk zijn. Afhankelijk van de alternatieve hypothese worden drie gevallen onderscheiden bij het toetsen:

a. *Rechtseenzijdige toetsing*, indien het interessegebied rechts ligt. Het zijn de hypothesen waarin steeds een $>$-teken voorkomt (zie fig. 9.2a).

b. *Linkseenzijdige toetsing*, indien het interessegebied aan de linkerkant ligt. Het betreft hypothesen waarin steeds een $<$-teken voorkomt. (zie fig. 9.2b).

c. *Tweezijdige toetsing*, indien het interessegebied aan weerszijden ligt. Dit is het geval in de eerste H_1 veronderstelling in de voorbeelden 3, 5 en 6 (zie fig. 9.2c). In de formulering komt een $\neq$-teken voor.

Dus afhankelijk van de probleemstelling onderscheiden we eenzijdige (links of rechts) en tweezijdige toetsen.



a. Rechteroverschrijdingskans b. linkeroverschrijdingskans c. Tweezijdige overschrijdingskans

Fig. 9.2 De drie verschillende overschrijdingskansen α

We hebben al gezien dat de onbetrouwbaarheid(sdrempel) of kans op een fout van de eerste soort aangegeven wordt door α. De waarde $1 - \alpha$ wordt de *betrouwbaarheid* genoemd. Daarnaast kennen we de kans op een fout van de tweede soort, aangeduid met β. De waarde $1 - \beta$ wordt het *onderscheidingsvermogen* van de toets genoemd.
Dit kan als volgt schematisch weer worden gegeven.

		werkelijkheid	
		H_0 juist	H_1 juist
Beslissing	H_0 verwerpen	α (= kans op fout van de eerste soort)	$1 - \beta$ (= onderscheidingsvermogen)
	H_0 niet verwerpen	$1 - \alpha$ (= betrouwbaarheid)	β (= kans op fout van de tweede soort)

Fig. 9.3 Beslissingsschema

9.2.3 Een uitgewerkt voorbeeld

De zojuist gegeven toetsingsprocedure zullen we aan de hand van voorbeeld 1 nalopen. Hierbij moesten we toetsen of behandeling met een chloorhoudende stof de slijtvastheid van rubber verhoogt. We hebben metingen verricht aan een aantal behandelde monsters en onbehandelde monsters en bepaalden de waarde van het verschil V in slijtvastheid van behandeld en onbehandeld rubber (zie de tabel bij voorbeeld 1).

1. *Probleemstelling*
   Er moet worden nagegaan of chloorbehandeling een positieve invloed heeft op de slijtweerstand.
2. *Opstellen van de hypothesen*
   H_0: $\mu_V = \mu_0 = 0$
   H_1: $\mu_V > 0$ (We zijn alleen geïnteresseerd in een positieve invloed.)
3. *Keuze van de onbetrouwbaarheid*
   $\alpha = 0{,}05$ (eenzijdig). Dat de toets eenzijdig is kunnen we afleiden uit procedurepunt 2 bij de alternatieve hypothese H_1.
4. *Keuze van de toetsingsvariabele*
   Als toetsingsvariabele nemen we het *gemiddelde* verschil in slijtweerstand ($\bar{V}$ met waarde $\bar{v}$), tussen behandeld rubber en onbehandeld rubber. Uit hoofdstuk 7 weten we dat $\bar{V}$, onder de veronderstelling dat de nulhypothese waar is, een normale verdeling volgt met $\mu_{\bar{V}} = \mu_V = 0$ en $\sigma_{\bar{V}} = \dfrac{\sigma_V}{\sqrt{n}}$. Om de toets daadwerkelijk te kunnen uitvoeren, hebben we de standaardafwijking nodig van de verschilvariabele V. Voor het gemak veronderstellen we de standaardafwijking van de gemiddelde verschillen bekend en onafhankelijk van de meetprocedure (later zullen we zien hoe we in werkelijkheid met deze standaardafwijking moeten omgaan). Neem aan dat $\sigma_V = 1{,}126$. De steekproefgrootte was 10 (zie voorbeeld 1). Hiermee is de verdeling van $\bar{V}$ bekend.
5. *Bepaling van de waarde van de toetsingsvariabele*
   Op basis van de steekproef vonden we $\bar{v} = 1{,}27$ (het gemiddelde verschil tussen onbehandeld en behandeld).
6. *Bepaling van de kritieke waarde*
   De toetsingsvariabele $\bar{V}$ volgt onder de nulhypothese een normale verdeling, met:
   $\mu_{\bar{V}} = 0$ en $\sigma_{\bar{V}} = 1{,}126$.
   We onderzoeken nu wat de kans is, dat $\bar{V} > 1{,}27$ in deze normale verdeling. Om deze kans te berekenen, gebruiken we de standaardnormale verdeling (u-verdeling). Daarvoor is de transformatie

$$U = \frac{\bar{V} - \mu_{\bar{V}}}{\sigma_{\bar{V}}} \tag{9.1}$$

   nodig. Bij de gevonden waarde voor $\bar{V}$ ($\bar{v} = 1{,}27$) hoort een u-waarde

$$u = \frac{1{,}27 - 0}{\frac{1{,}126}{\sqrt{10}}} = 3{,}57$$

De rechteroverschrijdingskans van 3,57 opzoeken in tabel B1 van de standaardnormale verdeling levert:

$P(U > 3{,}57) < 0{,}0002$

7. *Statistische conclusie*

   De overschrijdingskans van de toetsingsvariabele is kleiner dan $\alpha = 0{,}05$ (rechtseenzijdig). Conclusie: H_0 wordt verworpen ten gunste van H_1.

8. *Vertaling*

   Op grond van de steekproef van 10 strookjes rubber kan men concluderen dat chloorbehandeling een positieve invloed heeft op de slijtweerstand van rubber, met een onbetrouwbaarheid van maximaal 5%.

9.2.4 De samenhang tussen de constructie van betrouwbaarheidsintervallen en het toetsen van hypothesen

In hoofdstuk 8 hebben we met behulp van kansverdelingen betrouwbaarheidsintervallen voor de parameters van populaties geconstrueerd. Er is een relatie tussen deze betrouwbaarheidsintervallen en het toetsen van hypothesen ten aanzien van deze parameters. Het $(1-\alpha)$-betrouwbaarheidsinterval bevat alle waarden van de te onderzoeken toetsingsvariabele, waarbij de nulhypothese niet wordt verworpen (bij toetsing met een onbetrouwbaarheidsdrempel α). De grenswaarden van het betrouwbaarheidsinterval komen overeen met de zogenaamde 'kritieke' waarden bij het toetsen. Deze *kritieke waarden* vormen dus de grenswaarden, waarbij de nulhypothese nog net niet wordt verworpen. Waarden van de toetsingsvariabele die kleiner respectievelijk groter zijn dan deze kritieke waarden leiden tot het verwerpen van de nulhypothese. Dit leidt tot een iets andere opzet van de toetsingsprocedure.

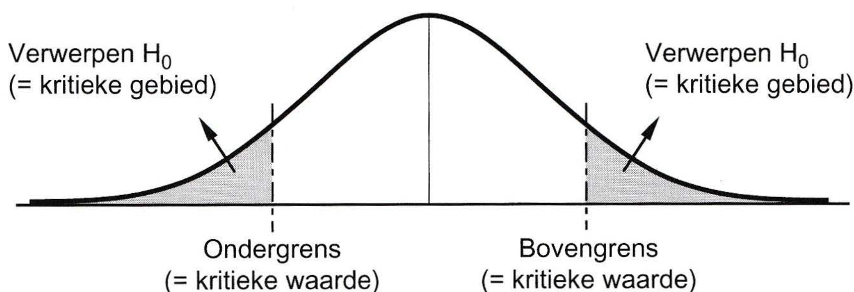


Fig. 9.4 Samenhang grenzen betrouwbaarheidsinterval en toetsen van hypothesen

Voorbeeld 7

Noem de opnameduur van een patiënt in een bepaald ziekenhuis X. De directie van het ziekenhuis zegt dat de gemiddelde opnameduur 5 dagen is. Uit een steekproef van

36 aselect gekozen patiënten zijn de volgende resultaten berekend over de opnameduur: $\bar{x} = 6{,}2$ en $s = 5{,}2$.
De vraag is nu: stemt dit resultaat overeen met de uitspraak van het ziekenhuis? (Toets met $\alpha = 0{,}05$.)

Oplossing

We zullen dit probleem oplossen met de vernieuwde opzet van een toetsingsprocedure.

1. Toetsingsprocedure: onderzocht moet worden of de waarde van de gemiddelde opnameduur in het ziekenhuis 5 dagen is.
2. H_0: $\mu_X = 5$ en H_1: $\mu_X \neq 5$.
3. $\alpha = 0{,}05$ (tweezijdig).
4. Het steekproefgemiddelde $\bar{X}$ is nu *niet* normaal verdeeld omdat de standaardafwijking onbekend is. Toetsingsvariabele is nu de variabele

$$T = \frac{\bar{X} - \mu_{\bar{X}}}{S_{\bar{X}}} = \frac{\bar{X} - \mu_{\bar{X}}}{\frac{S_X}{\sqrt{n}}} \tag{9.2}$$

   Hierin is $\mu_{\bar{X}} = \mu_X$ en S_X is de standaardafwijking van X. De waarde van $\frac{S_X}{\sqrt{n}}$ wordt vaak *standaardfout* genoemd.
   Deze variabele T volgt onder H_0 een t-verdeling met $\nu = n - 1$ vrijheidsgraden (verklaring: zie paragraaf 8.3.2).
5. Waarde van de toetsingsvariabele T? Om deze te kunnen berekenen, hebben we de steekproefresultaten nodig. Deze zijn $\bar{x} = 6{,}2$ en $s = 5{,}2$. Voor n kunnen we nemen $n = 36$.
6. Het vervolg van de toetsingprocedure verloopt nu op de vernieuwde manier (*a*) en op de 'oorspronkelijke' manier (*b*) als volgt.

   a. *Werken met kritieke waarden*
   (= grenzen van de tweezijdige betrouwbaarheidsinterval). De onbetrouwbaarheid α wordt naar beide zijden gelijk (= 0,025) verdeeld.
   Daar σ onbekend is, nemen we de waarde t van de t-verdeling met $\nu = n-1 = 35$ vrijheidsgraden en $\beta = 1 - \alpha = 0{,}95$ (tweezijdig). In de t-tabel (B5) vinden we (bij $\frac{1}{2}\alpha = 0{,}025$) een t-waarde $t = 2{,}03$.
   De kritieke waarden of grenswaarden van het betrouwbaarheidsinterval worden: $k_{1,2} = \bar{x} \pm t_{35}(\frac{1}{2}\alpha) \cdot \frac{s}{\sqrt{n}} = 5 \pm 2{,}03 \times \frac{5{,}2}{\sqrt{36}}$. Dus $k_1 = 3{,}24$ en $k_2 = 6{,}76$.
   Het 95%-betrouwbaarheidsinterval is: $3{,}24 \leq \mu \leq 6{,}76$.
   Het gebied waarbij de nulhypothese wordt verworpen (= kritiek gebied), wordt begrensd door de waarden: 3,24 en 6,76. In het gebied dat ligt tussen 3,24 en 6,76 wordt H_0 daarom niet verworpen.

De toetsingsvariabele $T = 6{,}2$ ligt daadwerkelijk tussen $k_1 = 3{,}24$ en $k_2 = 6{,}76$. De toetsingsvariabele ligt dus *niet* in het kritieke gebied. Conclusie: H_0 wordt niet verworpen.

b. Procedure met behulp van overschrijdingskansen

Iets anders verloopt de procedure als we eerst de overschrijdingskansen berekenen.

We bepalen de overschrijdingskans van T met waarde $\dfrac{6{,}2 - 5}{\frac{5{,}2}{\sqrt{36}}} = 1{,}38$ (onder de aanname dat H_0 waar is, dus $\mu_X = \mu_{\bar{X}} = 5$). De toetsingsvariabele is een trekking uit een t-verdeling met $\nu = 35$.

De overschrijdingskans wordt opgezocht in de t-tabel bij $\nu = 35$. Uit deze tabel (B5) valt op te maken dat:

$0{,}05 < P(T_{\nu=35} > 1{,}38) < 0{,}10$ (rechter kritieke waarden, dus eenzijdig).

Bij tweezijdige toetsing betekent dit (vermenigvuldiging van linker- en rechterlid met 2):

$0{,}10 < P(T_{\nu=35} > 1{,}38) < 0{,}20$.

De overschrijdingskans van T is in beide gevallen groter dan $\alpha = 0{,}05$. H_0 wordt daarom niet verworpen.

Op grond van dit onderzoek kan niet worden vastgesteld dat de gemiddelde opnametijd afwijkt van 5 dagen. ($\alpha = 0{,}05$).

9.3 Het toetsen met betrekking tot gemiddelden en spreidingen (de u-toets, t-toets en χ^2-toets)

In de vorige paragraaf zijn de begrippen en de principes van het toetsen, als ook de toepassingsprocedure ter sprake gekomen. In deze paragraaf bespreken we een aantal veelgebruikte 'klassieke' toetsen.

Deze toetsen zijn steeds gebaseerd op de veronderstelling dat de meetwaarden (bij benadering) normaal verdeeld zijn.

9.3.1 Toets voor een populatiegemiddelde waarbij σ bekend is (u-toets)

In de vorige paragraaf zijn de u-toets en de t-toets al even naar aanleiding van voorbeelden aan de orde geweest. We gaan nu deze toetsen nog eens formeel en algemeen bekijken.

Als we een toets willen uitvoeren ten aanzien van het gemiddelde μ_0 van een normaal verdeelde populatie (met kenmerk = variabele X), maken we gebruik van het steekproefgemiddelde $\bar{X}$ van een aselecte steekproef, getrokken uit die populatie. We gebruiken de waarde van het steekproefgemiddelde $\bar{x}$ om een conclusie te trekken over populatiegemiddelde.

Bij tweezijdige toetsing wordt H_0: $\mu = \mu_0$ (normwaarde) getoetst tegen H_1: $\mu \neq \mu_0$.

Neem aan dat de standaardafwijking van de populatie, waaruit de steekproef afkomstig is, bekend is (σ).

De toetsingsprocedure (op de oorspronkelijke manier, dus zonder de kritieke waarden te berekenen) verloopt als volgt.

1. *Vraagstelling*
   De vraag is of het gemiddelde van een bepaalde populatie (μ) gelijk is aan een gespecificeerde normwaarde (μ_0).
2. *Het opstellen van de hypothesen*
   H_0: $\mu = \mu_0$ en H_1: $\mu \neq \mu_0$
3. *Het vaststellen van de grootte van* α
   Voor α wordt meestal $\alpha = 0{,}05$ genomen. Vanuit de alternatieve hypothese wordt vastgesteld of α eenzijdig of tweezijdig moet worden genomen.
4. *Het vaststellen van de toetsingsvariabele*
   De toetsingsvariabele $\bar{X}$ is gebaseerd op het steekproefgemiddelde $\bar{x}$, dat onder aanname van H_0 een normale verdeling volgt. Aan het eind van hoofdstuk 7 hebben we gezien dat het steekproefgemiddelde $\bar{X}$ normaal verdeeld is met $\mu_{\bar{X}} = \mu_0$ en $\sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{n}}$.
   De verdeling van $\bar{X}$ is dus bekend omdat σ bekend is.
5. *Het berekenen van de toetsingsvariabele*
   Bepaal de waarde van de toetsingsvariabele $\bar{X}$, dus bereken $\bar{x}$.
6. *Het bepalen van de overschrijdingskans van de toetsingsvariabele*
   Bepaal de overschrijdingskans van $\bar{X}$, met behulp van $\bar{x}$.
   Bij rechtseenzijdig toetsen (als $\bar{x} > \mu_0$):

$$P(\bar{X} > \bar{x}) = P\left(U > \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right) \tag{9.3}$$

   Indien uit de alternatieve hypothese blijkt dat toetsing tweezijdig moet worden uitgevoerd, vermenigvuldig de gevonden overschrijdingskans dan met 2.
7. *Statistische conclusie*
   Vergelijk de gevonden overschrijdingskans van $\bar{X}$ met α en verwerp H_0 als deze overschrijdingskans kleiner is dan α.
8. *Rapportage*
   De vertaling naar de gestelde vraag in stap 1.
   Indien de alternatieve hypothese zodanig is dat er een eenzijdige toetsing moet worden uitgevoerd, verandert er in principe alleen in de procedurestappen 6 en 7 iets. We zijn dan maar in 'één kant' geïnteresseerd en de overschrijdingskans uit stap 6 wordt niet met 2 vermenigvuldigd.

Voorbeeld 8

Er wordt een onderzoek gedaan naar het afwijkende bestedingspatroon van een bepaalde provincie ten opzichte van het landelijk bestedingspatroon. Gekeken wordt naar de besteding per huishouden per week aan voeding. De gemiddelde landelijke besteding aan

voeding per huishouding is 158 euro per week, met een variantie van 900 euro^2. Een aselecte steekproef van $n = 100$ in de provincie geeft een gemiddelde besteding aan voeding van 168 euro per week. In een regionale krant wordt nu gesteld dat de besteding aan voeding in deze bepaalde provincie gemiddeld hoger ligt dan het landelijk gemiddelde. Kan op grond van de steekproef deze bewering onderschreven worden?

Oplossing

We doorlopen de genoemde toetsingsprocedure.

1. De vraag is of de gemiddelde besteding (per week) aan voeding in een bepaalde provincie hoger is dan het landelijk gemiddelde.

2. H_0: $\mu = 158$ en H_1: $\mu > 158$.

3. $\alpha = 0{,}05$ (eenzijdig).

4. $\bar{X}$ volgt onder de nulhypothese een (standaard-normale) u-verdeling ($\sigma^2 = 900$, dus is σ bekend).

5. De waarde van $\bar{X}$ is berekend: $\bar{x} = 168$.

6. $\sigma^2_{\bar{X}} = \dfrac{\sigma^2}{n} = \frac{900}{100} = 9$, dus $\sigma_{\bar{X}} = \sqrt{9} = 3$.

$$P(\bar{X} > 168) = P(U > \frac{168 - 158}{3}) = P(U > 3{,}33) = 0{,}0004.$$

7. De overschrijdingskans van $\bar{x}$ is veel kleiner dan $\alpha = 0{,}05$. De nulhypothese wordt daarom verworpen ten gunste van de alternatieve hypothese.

8. De besteding aan voeding in deze provincie is, zoals de krant vermeldde, inderdaad hoger dan het landelijk gemiddelde.

Opdracht

Bepaal het 'kritieke gebied' door die waarde k van $\bar{X}$ te berekenen waar vanaf de nulhypothese in twijfel getrokken moet worden (in dit geval de kleinste waarde van k waarvoor geldt $P(\bar{X} > k) \leq 0,05$).

Het onderscheidingsvermogen

Naar aanleiding van het laatste voorbeeld kunnen we ook iets zeggen over het *onderscheidingsvermogen* $(1 - \beta)$. Indien H_0 niet wordt verworpen, wil dit nog niet zeggen dat H_0 inderdaad ook juist is.

In figuur 9.5 is deze situatie voor het laatste voorbeeld in beeld gebracht.

We willen de fout van de tweede soort berekenen voor een van de alternatieven, bijvoorbeeld voor het gemiddelde: $\mu = 171$.

Indien H_1 waar is, volgt het steekproefgemiddelde $\bar{X}$ in dat geval een normale verdeling met gemiddelde $\mu_{\bar{X}} = 171$ en variantie $\sigma^2_{\bar{X}} = \frac{900}{100} = 9$.

De kritieke waarde (k) vinden we met behulp van de onbetrouwbaarheid $\alpha = 0{,}05$, onder de voorwaarde dat de nulhypothese waar is.

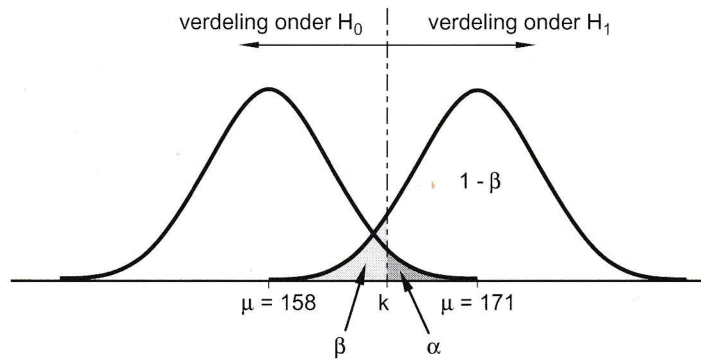


Fig. 9.5 Kans op een fout van de tweede soort

Er geldt: $u(0{,}05) = \dfrac{k - 158}{\sqrt{9}}$, met $u(0{,}05) = 1{,}65$ (tabel B1) dus: $k = 1{,}65 \times \sqrt{9} + 158 = 4{,}95 + 158 = 162{,}94$.

Voor het bepalen van de kans op een fout van de tweede soort (β), bepalen we de kans op een uitkomst kleiner dan de kritieke waarde $k = 162{,}94$, onder de aanname dat H_1 ($\mu = 171$) waar is.

Onder aanname van H_1 geldt:

$$\beta = P\left(U < \frac{162{,}94 - 171}{\sqrt{9}}\right) = P(U < -2{,}69) = P(U > 2{,}69) = 0{,}0036$$

We zien dat de β-fout erg klein is. Dit is te danken aan twee omstandigheden, namelijk dat:

- n vrij groot is, waardoor de standaardafwijking van $\bar{X}$ ($= \sigma_{\bar{X}}$) klein wordt;
- de gemiddelden van de nulhypothese en de alternatieve hypothese tamelijk ver uit elkaar liggen.

In zijn algemeenheid kunnen we opmerken dat we de β-fout kleiner kunnen krijgen, door de steekproefomvang groter te nemen. Daardoor vergroten we ook het onderscheidingsvermogen ($1 - \beta$) bij een gegeven α.

In dit voorbeeld toetsten we met de u-toets (σ is bekend) of een gemiddelde μ ongelijk is aan een gespecificeerde μ_0.

Onder H_0 volgt de variabele $U = \dfrac{\bar{X} - \mu}{\sigma_{\bar{X}}}$ een standaardnormale verdeling ($N(0, 1)$), met $\sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{n}}$ (n is de steekproefgrootte).

Indien H_0: $\mu = \mu_0$ waar is, is de kans om H_0 ten onrechte te verwerpen maximaal gelijk aan α.

Stel nu: $\mu = \mu_1 \neq \mu_0$. In dit geval is de kans om H_0 te verwerpen groter dan α en des te groter naarmate μ_1 verder van μ_0 af ligt. Ook geldt: de kans dat H_0 in dat geval (ten onrechte) aanvaard wordt is β, dus de kans om H_0 (terecht) te verwerpen is $1 - \beta$. We kunnen voor verschillende waarden van μ de kans berekenen op het verwerpen van H_0 en

dit grafisch uitzetten als functie van μ (zie figuur 9.6). We noemen de ontstane curve de *kromme van het onderscheidingsvermogen (OC-curve).*

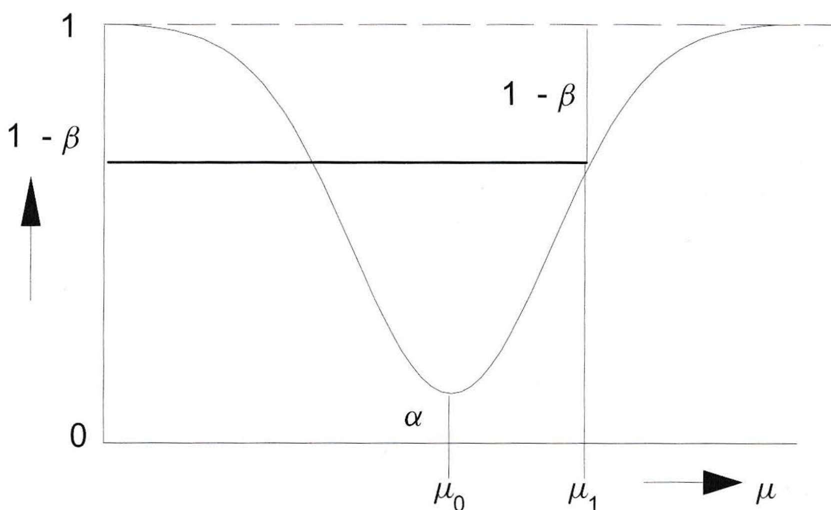


Fig. 9.6 Kromme van het onderscheidingsvermogen (OC-curve)

9.3.2 Toets voor een populatiegemiddelde met onbekende σ^2 (t-toets)

In de meeste gevallen zal de variantie σ^2 niet bekend zijn. We kunnen dan de u-toets niet gebruiken, maar zijn aangewezen op de zogenaamde t-toets, waarbij we in plaats van u-waarden, de t-waarden gebruiken van de t-verdeling van Student.

Uit hoofdstuk 8 weten we: indien we een steekproef hebben uit een populatie met populatiegemiddelde μ en onbekende variantie, volgt het steekproefgemiddelde, onder H_0: $\mu = \mu_0$, een t-verdeling met $\nu = n - 1$ vrijheidsgraden.

De overschrijdingskans van $\bar{X}$ wordt bepaald met de toetsingsvariabele: $T = \dfrac{\bar{X} - \mu_o}{\frac{S}{\sqrt{n}}}$.

De toetsingsprocedure is, op een aantal kleine onderdelen na, dezelfde als die uit de vorige paragraaf.

Aan de hand van voorbeeld 9 zullen we de toetsing weer stap voor stap uitvoeren.

Voorbeeld 9

Een fabrikant wil weten of het zoutgehalte in een partij mosterd hoger is dan 16%. Hij neemt daartoe een aselecte steekproef van $n = 5$ uit de partij en wenst bij toetsing een onbetrouwbaarheid $\alpha = 0{,}05$. De uitkomsten van de steekproef zijn: 15,7 – 16,3 – 16,5 – 15,9 en 16,3

Oplossing

1. Is het zoutgehalte van mosterd hoger dan 16%?

2. H_0: $\mu = 16\%$ en H_1: $\mu > 16\%$
3. $\alpha = 0{,}05$ (eenzijdig).
4. Als toetsingsvariabele nemen we

$$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \tag{9.4}$$

die onder de nulhypothese een t-verdeling volgt met $\nu = n - 1$ vrijheidsgraden.

5. Uit de steekproef bepalen we de waarde van $\bar{X}$ $(= \bar{x})$ en S $(= s)$ en vinden: $\bar{x} = 16{,}14$ en $s = 0{,}329$.
6. We bepalen de overschrijdingskans van T door de bijbehorende waarde te berekenen onder de voorwaarde dat H_0 waar is ($\mu = 16$):

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{16{,}14 - 16}{\frac{0{,}329}{\sqrt{5}}} = 0{,}95$$

In de t-tabel vinden we bij $\nu = 5 - 1 = 4$ een overschrijdingskans die groter is dan 10% (eenzijdig).

7. De gevonden overschrijdingskans is groter dan $\alpha = 0{,}05$ en dus wordt H_0 niet verworpen.
8. Op grond van dit onderzoek is niet aangetoond dat het zoutgehalte in mosterd hoger is dan 16%.

Opdracht

Bepaal de kritieke waarde voor $\bar{X}$, dat wil zeggen de kleinste waarde k waarvoor geldt: $P(\bar{X} > K) \leq 0{,}05$, zodat bij een steekproefgemiddelde $\geq k$ de nulhypothese verworpen dient te worden.

9.3.3 Toets voor een fractie

Het toetsen van een fractie p van een binomiale verdeling lichten we toe aan de hand van het volgende voorbeeld.

Voorbeeld 10

Voor een oude wijk van gemeentewoningen is enige tijd geleden een renovatieplan ontworpen waarmee, blijkens een toen gehouden onderzoek een fractie 0,75 van de hoofdbewoners in principe kon instemmen. Inmiddels zijn enige financiële aspecten duidelijker naar voren gekomen, onder andere de noodzakelijke huurverhoging voor de gerenoveerde woningen. Een actiegroep is nu van mening dat de fractie p van instemmers thans beduidend lager is dan 0,75 en acht een nieuw onderzoek noodzakelijk. De gemeente vindt een volledig nieuw onderzoek te kostbaar en wenst de hypothese $p = 0{,}75$ te toetsen tegenover de alternatieve hypothese $p < 0{,}75$ op grond van een aselecte steekproef

van 100 hoofdbewoners. Het aantal instemmers in de steekproef blijkt gelijk te zijn aan 70. Blijkt nu, aan de hand van dit onderzoek, dat de fractie instemmers kleiner is geworden? We nemen aan dat de omvang van de populatie oneindig groot is ten opzichte van de steekproefomvang.

Oplossing

De toetsingsprocedure verloopt als volgt:

1. Is de fractie instemmers kleiner dan $p = 0{,}75$?
2. H_0: $p = 0{,}75$ en H_1: $p < 0{,}75$.
3. $\alpha = 0{,}05$ (eenzijdig).
4. Als toetsingsvariabele nemen we K = het aantal instemmers in de steekproef. K volgt, onder aanname van H_0 een binomiale verdeling met $n = 100$ en $p = 0{,}75$. Daar de steekproefomvang ($n = 100$) groter is dan $9\frac{1-p}{p} = 9 \times \frac{0{,}75}{0{,}25} = 27$, mogen we deze binomiale verdeling (zoals aangetoond in hoofdstuk 6) benaderen door een normale verdeling met $\mu = np = 100 \times 0{,}75 = 75$ en $\sigma = \sqrt{np(1-p)} = \sqrt{100 \times 0{,}75 \times 0{,}25} = 4{,}33$.
5. In de steekproef hebben we 70 instemmers gevonden.
6. We bepalen nu de overschrijdingskans van $K = 70$, indien de nulhypothese waar is (let op de continuïteitscorrectie):
   $P(K \leq 70 | H_0) = P(U < \frac{70{,}5-75}{4{,}33}) = P(U < -1{,}04) = 0{,}1492$ (eenzijdig).
7. De gevonden overschrijdingskans is groter dan $\alpha = 0{,}05$ en dus wordt H_0 niet verworpen.
8. Op grond van dit onderzoek mag geconcludeerd worden dat $p = 0{,}75$ nog steeds geldt, althans dat het resultaat van de steekproef geen reden vormt om hiervan af te wijken.

Opdracht

Bereken de kritieke waarde voor p, dus de grootste waarde k waarvoor geldt:
$P(p < k) \leq 0{,}05$.

In dit voorbeeld hebben we kunnen zien dat het toetsen van een hypothese die geldt voor een fractie, kan geschieden met behulp van de normale verdeling. Dit was mogelijk omdat de steekproef groot genoeg was.

9.3.4 Het toetsen van een variantie; toetsing met behulp van de χ^2-verdeling

Naast het toetsen van een (hypothese ten aanzien van) een gemiddelde of een fractie is er in de praktijk ook vaak behoefte aan het toetsen van een variantie. Met behulp van de χ^2-verdeling (zie ook paragraaf 8.4) is een dergelijke toetsing mogelijk.

De toetsingsvariabele is de steekproefvariantie. Indien de meetwaarden (bij benadering) normaal verdeeld zijn, volgt de variabele:

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \qquad (9.5)$$

een χ^2-verdeling, met $\nu = n - 1$ vrijheidsgraden. De toetsing zullen we uitvoeren aan de hand van het volgende voorbeeld.

Voorbeeld 11

De standaardafwijking van de (normaal verdeelde) diameters van nylon kogeltjes voor een bepaald soort kogellagers mag overeenkomstig de geldende kwaliteitsnorm niet meer bedragen dan 0,3 mm. Bij een controle vond men in een steekproef van $n = 25$ kogeltjes een standaardafwijking $s = 0{,}4$ mm. Volgt hieruit dat de dagproductie waaruit deze steekproef afkomstig is wat de spreiding van de diameters van de kogels betreft niet aan de gestelde eis voldoet?

Oplossing

De toetsingsprocedure verloopt als volgt.

1. Is de standaardafwijking van de kogeldiameters D in de bedoelde dagproductie groter dan $\sigma = 0{,}3$ mm?
2. H_0: $\sigma_D^2 = 0{,}09$ en H_1: $\sigma_D^2 > 0{,}09$.
3. $\alpha = 0{,}05$ (eenzijdig).
4. Als we de variantie van de diameters in een steekproef van n kogels uit een dagproductie aanduiden met S_D^2, volgt de kansvariabele $\chi^2 = \dfrac{(n-1)S_D^2}{\sigma_D^2}$ een χ^2-verdeling met $\nu = n - 1$.
5. De variantie in de steekproef van 25 waarnemingen bedraagt 0,16.
6. De overschrijdingskans van $c = \dfrac{(n-1)s_D^2}{\sigma_D^2} = \frac{24\times 0{,}16}{0{,}09} = 42{,}67$ wordt opgezocht in de χ^2-tabel (tabel B6) bij $\nu = 25 - 1 = 24$ vrijheidsgraden. We vinden een overschrijdingskans die kleiner is dan 0,01 (eenzijdig).
7. De gevonden overschrijdingskans is kleiner dan $\alpha = 0{,}05$. De nulhypothese wordt daarom verworpen ten gunste van de alternatieve hypothese.
8. De spreiding van de diameters in de dagproductie wijkt af van de norm $\sigma_D = 0{,}3$ mm.

Opdracht

Bepaal het kritieke gebied voor S_D^2.

Indien de steekproefomvang voldoende groot ($n > 30$) is, mag men het geheel ook benaderen door een normale verdeling met $\mu = \nu$ en $\sigma = \sqrt{2\nu}$ (zie hoofdstuk 8, bij chikwadraatverdeling).

Voorbeeld 12

We gaan uit van het vorige voorbeeld, maar nemen nu een steekproef van 65 kogels in plaats van 25. De standaardafwijking in de diameters in deze steekproef is $s_D = 0{,}4$ mm.

Oplossing

Voor de toetsingsprocedure verandert er alleen iets in de bepaling van de overschrijdingskans van χ^2. Onder H_0 volgt $\chi^2 = \dfrac{(n-1)S_D^2}{\sigma_D^2}$ een normale verdeling met $\mu_{\chi^2} = 65$ en $\sigma_{\chi^2} = \sqrt{2 \times 65} = 11{,}40$. De overschrijdingskans van χ^2 berekenen we als volgt:
De waarde c van χ^2 op basis van de steekproef is $\dfrac{(n-1)s_D^2}{\sigma_D^2} = \frac{64 \times 0{,}16}{0{,}09} = 113{,}78$.
$P(\chi^2 > 113{,}78) = P(U > \frac{113{,}78-65}{11{,}40}) = P(U > 4{,}27) \ll 0{,}0002$.
Deze kans is veel kleiner dan $\alpha = 0{,}05$ en dus wordt H_0 verworpen ten gunste van H_1.

9.4 Vergelijkings- of verschiltoetsen

In voorbeeld 1 vergeleken we de slijtvastheid van rubber na en voor behandeling met een chemische stof. Om deze vergelijking te kunnen maken, hebben we het verschil bepaald van de slijtvastheid na en voor de bewerking. Met behulp van de verschilvariabele V hebben we vervolgens een toets uitgevoerd, waarbij als toetsingsvariabele het gemiddelde verschil werd gehanteerd. We zagen hier al een voorbeeld van een *vergelijkingstoets* oftewel een *verschiltoets*. In deze paragraaf zullen we hier dieper op ingaan. Eerst zullen we twee verschiltoetsen laten zien ten aanzien van het gemiddelde verschil, daarna zullen we ook varianties en fracties met elkaar vergelijken.
De twee eerst te behandelen toetsen zijn uitsluitend toepasbaar voor gemiddelden van aselecte steekproeven uit (bij benadering) normale verdelingen. Wanneer niet aan deze eis voldaan is, zullen andere, zogenaamde verdelingsvrije methoden gebruikt moeten worden. Deze vallen buiten het kader van dit boek.
Aan de hand van het volgende voorbeeld zullen we zien dat we bij het toetsen van de gelijkheid van twee gemiddelden, twee situaties moeten onderscheiden: *gepaarde* en *niet-gepaarde* waarnemingen.

Voorbeeld 13

We willen de slijtweerstand van een bepaalde rubbersoort verbeteren door het rubber te behandelen met een chloorhoudende stof. Voor het onderzoek kunnen we uitgaan van twee soorten 'proefopzetten'.

Experiment 1

Uit een groot aantal stukken rubber kiezen we aselect 10 stukken. Elk van de 10 stukken verdelen we in tweeën. Het ene deel krijgt een behandeling met de chloorhoudende stof,

het andere deel blijft onbehandeld. Daarna wordt in willekeurige volgorde de slijtweerstand van de 20 stukken rubber bepaald.

Experiment 2
Uit een groot aantal stukken rubber kiezen we aselect 20 stukken. Daarna verdelen we de 20 stukken rubber aselect in twee groepen van 10 stukken. De ene groep van 10 stukken rubber wordt behandeld met de chloorhoudende stof, de andere groep blijft onbehandeld. Daarna wordt in willekeurige volgorde de slijtweerstand van de 20 stukken rubber bepaald.

In experiment 1 hebben we paren gevormd van steeds één stuk rubber. Van ieder paar wordt er aselect één behandeld en één blijft onbehandeld.
Bij experiment 2 hebben we 20 onafhankelijke stukken rubber, waaruit twee steekproeven zijn gevormd (wel en niet behandeld), die onafhankelijk van elkaar zijn. De twee steekproeven (wel en niet behandeld) in experiment 1 zijn niet onafhankelijk. De stukken rubber worden paarsgewijs vergeleken.
Het verschil in beide experimenten komt duidelijk naar voren bij de rekenprocedure en de toetsingsprocedure.
We zullen de twee experimenten uitvoeren aan de hand van dezelfde gegevens. Hierdoor komt dan de 'winst' van een opzet met gepaarde waarnemingen duidelijk naar voren. Opgemerkt dient echter te worden dat we in de praktijk niet kunnen kiezen welke rekenprocedure we kunnen nemen. Op grond van het experiment ligt de toetsingsprocedure vast.
In beide experimenten willen we onderzoeken of er verschil bestaat tussen behandelde en onbehandelde monsters. In tabel 9.2 zijn de benodigde gegevens weergegeven.
Bij het experiment met *gepaarde waarnemingen* bepalen we steeds het verschil in slijtweerstand van het uit één stuk rubber afkomstige paar, waarbij een deel behandeld is en het andere deel onbehandeld is gebleven.
In tabel 9.2 (laatste kolom) zien we dat het gemiddelde verschil tussen de 10 paren behandelde en onbehandelde stukken rubber gelijk is aan: $\bar{v} = 1{,}27$.
De variantie van de verschillen is: $s_V^2 = 1{,}269$.
De variantie van het gemiddeld verschil is:

$$s_{\bar{V}}^2 = \frac{s_V^2}{n} = \frac{1{,}269}{10} = 0{,}1269.$$

Bij de twee *onafhankelijke steekproeven* bepalen we eerst het gemiddelde en de variantie van de behandelde stukken rubber (steekproef X) en daarna het gemiddelde en de variantie van de onbehandelde stukken rubber (steekproef Y). Het gemiddelde verschil tussen behandelde en onbehandelde monsters heeft een waarde: $\bar{v} = \bar{x} - \bar{y} = 12{,}87 - 11{,}60 = 1{,}27$.

Tabel 9.2 Slijtweerstand van rubbermonsters

paar nr.	slijtweerstand behandeld (X) x	onbehandeld (Y) y	verschil ($V = X - Y$) $v = x - y$
1	14,7	12,1	2,6
2	14,0	10,9	3,1
3	12,9	13,1	-0,2
4	16,2	14,5	1,7
5	10,2	9,6	0,6
6	12,4	11,2	1,2
7	12,0	9,8	2,2
8	14,8	13,7	1,1
9	11,8	12,0	-0,2
10	9,7	9,1	0,6
	$\bar{x} = 12{,}87$	$\bar{y} = 11{,}60$	$\bar{v} = 1{,}27$
	$s_X^2 = 4{,}305$	$s_Y^2 = 3{,}291$	$s_V^2 = 1{,}269$

De variantie van het gemiddelde verschil $s_{\bar{V}}^2$ vinden we als volgt (bij het verschil van twee variabelen moeten we de varianties optellen):

$$s_{\bar{V}}^2 = s_{\bar{X}}^2 + s_{\bar{Y}}^2 = \frac{s_X^2}{n} + \frac{s_Y^2}{n} = \frac{4{,}305}{10} + \frac{3{,}291}{10} = 0{,}4305 + 0{,}3291 = 0{,}7596$$

We zien dat het gemiddelde verschil in beide proefopzetten gelijk is aan $\bar{v} = 1{,}27$. De variantie van de gemiddelde verschillen wijken echter nogal van elkaar af (0,1269 ten opzichte van 0,7596). Toch zijn beide varianties schattingen van dezelfde σ^2.
Het grote verschil tussen de eerste en de tweede schatting van σ^2 wordt verklaard, doordat de tamelijk grote verschillen *tussen* de stukken rubber bij experiment 1 worden geëlimineerd en in experiment 2 volledig in de berekening worden meegenomen.
We kunnen dit als volgt zien:
Stel we tellen bij paar nr.8, zowel bij de x- als bij de y-waarde 20,0 op, hierdoor verandert het verschil tussen x en y bij paar nr.8 niet; dus ook niet de schatting van σ^2. Bij de twee onafhankelijke steekproeven wordt de variantie van zowel (steekproef) X, als van (steekproef) Y groter en daarmee ook de schatting van σ^2.
Omdat wij een verschil tussen de behandelde en de onbehandelde stukken rubber willen aantonen, vormen de 'normale' verschillen tussen stukken rubber een storende factor, die we graag willen elimineren. Als de omstandigheden het toelaten, zullen we altijd uitgaan van een experiment met gepaarde waarnemingen. Hierbij moet direct worden opgemerkt dat dit niet altijd mogelijk is.

We zullen ons nu toeleggen op de toetsing van beide typen experimenten. We maken dus onderscheid tussen:

a. toetsing bij twee gepaarde steekproeven;

b. toetsing bij twee onafhankelijke steekproeven.

9.4.1 Toets voor het verschil van twee gemiddelden bij gepaarde waarnemingen

Deze toets is in principe gelijk aan de t-toets voor één populatiegemiddelde. We werken uitsluitend met het verschil V, binnen de paren van waarnemingen ($v_i = x_i - y_i$).
Bij toetsing van de nulhypothese H_0: $\mu_V = 0$ gaan we uit van de veronderstelling dat de verschillen $v_1, v_2, ...,$ onderling onafhankelijke, aselecte trekkingen zijn uit een normale verdeling met gemiddelde $\mu_V = 0$ en variantie σ^2. Als toetsingsvariabele gebruiken we $\bar{V}$, waarbij geldt dat onder aanname van H_0 de variabele $T = \dfrac{\bar{V} - \mu_{\bar{V}}}{S_{\bar{V}}}$ (met waarde: $t = \dfrac{\bar{v} - \mu_v}{s_{\bar{v}}} = \dfrac{\bar{v} - \mu_v}{\frac{s_v}{\sqrt{n}}}$) een t-verdeling volgt met $n - 1$ vrijheidsgraden, waarbij n het aantal paren is. We volgen de toetsingsprocedure voor de gegevens van de tabel (9.2).

1. Heeft een chloorbehandeling van rubber tot gevolg dat de slijtweerstand wordt verhoogd?

2. H_0: $\mu_V = 0$ en H_1: $\mu_V > 0$

3. $\alpha = 0{,}05$ (eenzijdig).

4. Toetsingsvariabele $\bar{V}$, die onder de nulhypothese een t-verdeling volgt met $n - 1$ vrijheidsgraden.

5. De gevonden waarde voor $\bar{v}$ is 1,27. De overschrijdingskans van $\bar{V}$ is voor deze waarde:

$$P(\bar{V} > 1{,}27) = P\left(T > \frac{1{,}27 - 0}{\sqrt{\frac{1{,}269}{10}}}\right) = P(T > 3{,}57).$$

6. In de t-tabel (B5) vinden we bij $\nu = 9$, dat $0{,}001 < P(T > 3{,}57) < 0{,}005$.

7. De gevonden overschrijdingskans is kleiner dan $\alpha = 0{,}05$. De nulhypothese wordt verworpen ten gunste van de alternatieve hypothese (onbetrouwbaarheid $\alpha = 0{,}05$).

8. Door de rubber met een chloorhoudende stof te behandelen, verbetert de slijtweerstand.

9.4.2 Toets voor het verschil van twee gemiddelden van twee onafhankelijke steekproeven

We gaan uit van twee normaal verdeelde populaties van de kansvariabelen X en Y met gemiddelde respectievelijk μ_X en μ_Y en varianties σ_X^2 en σ_Y^2. Uit elk van deze beide populaties wordt een aselecte steekproef getrokken van omvang n_X respectievelijk n_Y (de beide steekproeven hoeven dus niet even groot te zijn).
Voor de toetsing van de hypothese dat de twee populatiegemiddelden μ_X en μ_Y van elkaar verschillen, kunnen we de t-verdeling gebruiken, als aan de voorwaarde is voldaan dat

$\sigma_X^2 = \sigma_Y^2 = \sigma^2$. Indien niet aan deze voorwaarde is voldaan, is exacte toetsing zeer moeilijk.

Op de bekende wijze berekenen we de steekproefgemiddelden $\bar{X}$ en $\bar{Y}$ (waarden: $\bar{x}$ en $\bar{y}$) en tevens s_X^2 en s_Y^2, met respectievelijk $\nu_X = n_X - 1$ en $\nu_Y = n_Y - 1$ vrijheidsgraden. Als toetsingsvariabele wordt gebruikt: $V = \bar{X} - \bar{Y}$, waarbij onder de nulhypothese geldt dat de variabele

$$T = \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{S^2(\frac{1}{n_x} + \frac{1}{n_y})}} \tag{9.6}$$

een t-verdeling volgt met

$$\nu = (n_X - 1) + (n_Y - 1) = n_X + n_Y - 2 \tag{9.7}$$

vrijheidsgraden.

De variabele S^2 (onder het wortelteken) staat bekend als de *gepoolde steekproefvariantie.* De gepoolde steekproefvariantie wordt gebruikt als schatter van de populatievariantie σ^2. Deze gepoolde variantie is het *gewogen gemiddelde* van de twee afzonderlijke steekproefvarianties, met als weegfactoren het aantal vrijheidsgraden van de afzonderlijke steekproefvariantie (voor de bijbehorende formule, zie stap 4 in de navolgende toetsingsprocedure). Deze *pooling* mag alleen als zowel s_X^2 als s_Y^2 schatters zijn van dezelfde σ^2 (of de standaardafwijkingen van beide populaties daadwerkelijk gelijk zijn is te toetsen met behulp van de zogenaamde F-toets, zie de volgende paragraaf).

Voor de gegevens van het voorbeeld krijgen we de volgende procedure:

1. Heeft de chloorbehandeling van rubber tot gevolg dat de slijtweerstand wordt verhoogd?
2. H_0: $\mu_X = \mu_Y$ oftewel $\mu_X - \mu_Y = 0$ en H_1: $\mu_X > \mu_Y$.
3. $\alpha = 0{,}05$ (eenzijdig).
4. T heeft als waarde $t = \dfrac{(\bar{x} - \bar{y}) - 0}{\sqrt{s^2(\frac{1}{n_X} + \frac{1}{n_Y})}}$, met

$$s^2 = \frac{\nu_X s_X^2 + \nu_Y s_Y^2}{\nu_X + \nu_Y} \tag{9.8}$$

onder de voorwaarde dat s_X^2 en s_Y^2 schatters zijn van dezelfde σ^2.
T volgt onder H_0 een t-verdeling met $\nu = \nu_X + \nu_Y$ vrijheidsgraden.

5. $s^2 = \dfrac{9 \times 4{,}305 + 9 \times 3{,}291}{9 + 9} = 3{,}798$

$t = \dfrac{12{,}87 - 11{,}60}{\sqrt{3{,}798(\frac{1}{10} + \frac{1}{10})}} = 1{,}46$, met $\nu = 9 + 9 = 18$ vrijheidsgraden.

Tabel 9.3 Siliciumgehalte (%) van gietijzeren staven

partij 1	partij 2
11,3	7,7
9,2	7,4
8,6	8,9
8,7	8,5
1,.0	7,5
9,7	6,9
8,5	9,4
11,2	8,7
	6,4
	8,2
$\bar{x}_1 = 9.65$	$\bar{x}_2 = 7.96$
$s_1^2 = 1.254$	$s_2^2 = 0.889$

6. De overschrijdingskans van T bij een waarde $t = 1{,}46$ is te vinden met behulp van de t-tabel (B5) bij $\nu = 18$ vrijheidsgraden.
We vinden: $0{,}05 < P(T > 1{,}46) < 0{,}10$ (eenzijdig).

7. De overschrijdingskans van T is groter dan $\alpha = 0{,}05$. Conclusie: H_0 wordt niet verworpen.

8. Op grond van de gegevens uit de twee steekproeven, kan men niet concluderen dat een chloorbehandeling van rubber een vergroting van de slijtweerstand geeft.

We zien dat in de bewoordingen van stap 8 de formulering erg voorzichtig is, want zoals we al gezien hebben, is bij een andere proefopzet (gepaarde waarnemingen) wel een vergroting van de slijtweerstand aan te tonen.
Neem daarom de raad aan: 'is het mogelijk om gepaarde waarnemingen te vergelijken, doe het'.

Voorbeeld 14

Om het siliciumgehalte (in %) van twee partijen gietijzeren staven te vergelijken, worden uit beide partijen een aselecte steekproef getrokken en hiervan wordt het siliciumgehalte bepaald. De resultaten zijn weergegeven in tabel 9.3.

Oplossing

Voor de toetsing doorlopen we weer stapsgewijs de toetsingsprocedure.

1. Bestaat er verschil in siliciumgehalte van de partijen?

2. H_0: $\mu_1 = \mu_2$ oftewel $\mu_1 - \mu_2 = 0$ en H_1: $\mu_1 \neq \mu_2$.

3. $\alpha = 0{,}05$ (tweezijdig).
4. Als $\bar{X}_1$ en $\bar{X}_2$ de gemidddelden zijn van de beide steekproeven (met waarden $\bar{x}_1$en $\bar{x}_2$, volgt toetsingsvariabele $T = \bar{X}_1 - \bar{X}_2$, onder H_0 een t-verdeling, met $\nu = (n_1 - 1) + (n_2 - 1)$ vrijheidsgraden.
5. $\bar{x}_1 = 9{,}65$ en $s_1^2 = 1{,}254$.
   $\bar{x}_2 = 7{,}96$ en $s_2^2 = 0{,}889$.
   σ_1^2 en σ_2^2 zijn niet bekend, maar s_1^2 en s_2^2 verschillen echter weinig van elkaar (dit moeten we feitelijk toetsen; hierop komen we in de volgende paragraaf terug). We nemen aan dat σ_1^2 en σ_2^2 niet van elkaar verschillen en berekenen de gepoolde variantie s^2:
   $$s^2 = \frac{\nu_1 s_1^2 + \nu_2 s_2^2}{\nu_1 + \nu_2} = \frac{7 \times 1{,}254 + 9 \times 0{,}889}{7 + 9} = \tfrac{16{,}784}{16} = 1{,}049.$$
   $$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{s^2(\frac{1}{n_1} + \frac{1}{n_2})}} = \frac{9{,}65 - 7{,}96}{\sqrt{1{,}049(\frac{1}{8} + \frac{1}{10})}} = \tfrac{1{,}69}{0{,}486} = 3{,}48.$$
6. De overschrijdingskans van T bij waarde $t = 3{,}48$ berekenen we met behulp van de t-tabel met $\nu = \nu_1 + \nu_2 = 7 + 9 = 16$.
   We vinden: $0{,}001 < P(T > 3{,}48) < 0{,}005$ (eenzijdig) of $0{,}002 < P(T > 3{,}48) < 0{,}01$ (tweezijdig).
7. De overschrijdingskans van T is kleiner dan $\alpha = 0{,}05$. H_0 wordt verworpen ten gunste van H_1.
8. De conclusie luidt: 'de twee partijen gietijzeren staven hebben een verschillend siliciumgehalte'.

9.4.3 Het vergelijken van twee varianties (F-toets)

Om de gelijkheid van de varianties van twee normaal verdeelde populaties te toetsen, dienen we te beschikken over twee onafhankelijke aselecte steekproeven met omvang n_A respectievelijk n_B uit deze populaties. Van de beide steekproeven wordt de steekproefvariantie bepaald (S_A^2 met waarde s_A^2 en S_B^2 met waarde s_B^2). De toetsingsvariabele

$$F = \frac{S_A^2}{S_B^2} \tag{9.9}$$

met waarde $f = \frac{s_A^2}{s_B^2}$, volgt onder de nulhypothese H_0: $\sigma_A^2 = \sigma_B^2$, een zogenaamde F-verdeling (genoemd naar de statisticus R.A. Fisher (1924)) met $\nu_1 = n_A - 1$ en $\nu_2 = n_B - 1$ vrijheidsgraden.

Opmerking

De F-verdeling wordt volledig bepaald door het aantal vrijheidsgraden $\nu_1 = n_1 - 1$ van de variantie in de teller, respectievelijk $\nu_2 = n_2 - 1$ van de variantie in de noemer. Voor verschillende combinaties van ν_1 en ν_2 zijn de kritieke F-waarden getabelleerd in de

tabellen B7, B8 en B9). In deze tabellen zijn alleen rechter kritieke waarden opgenomen. Op de theorie van de F-verdeling zullen we in dit boek niet verder ingaan. We zullen deze verdeling uitsluitend gebruiken om te toetsen of twee varianties aan elkaar gelijk zijn.

Om H_0: $\sigma_A^2 = \sigma_B^2$ te toetsen, wordt, afhankelijk van de alternatieve hypothese H_1, de waarde f van de toetsingsvariabele F als volgt berekend:

a. H_1: $\sigma_A^2 < \sigma_B^2$, dan is $f = \frac{s_B^2}{s_A^2}$

De eenzijdige overschrijdingskans wordt afgelezen in de F-tabel bij $\nu_1 = n_B - 1$ en $\nu_2 = n_A - 1$.

b. H_1: $\sigma_A^2 > \sigma_B^2$, dan is $f = \frac{s_A^2}{s_B^2}$

De eenzijdige overschrijdingskans wordt afgelezen in de F-tabel bij $\nu_1 = n_A - 1$ en $\nu_2 = n_B - 1$.

c. H_1: $\sigma_A^2 \neq \sigma_B^2$, dan is $f = \frac{\text{grootste steekproefvariantie}}{\text{kleinste steekproefvariantie}}$

De eenzijdige overschrijdingskans wordt afgelezen in de F-tabel bij ν_1 en ν_2, met ν_1 = aantal vrijheidsgraden van de variantie in de teller en ν_2 = aantal vrijheidsgraden van de variantie in de noemer. Indien tweezijdig getoetst moet worden, wordt de gevonden overschrijdingskans met twee vermenigvuldigd.

Door consequent bovenstaande berekeningswijze toe te passen wordt bereikt dat uitsluitend te grote (significant groter dan 1) F-waarden tot verwerping van H_0 leiden.

Voorbeeld 15

In het geval van het laatste voorbeeld hebben we aangenomen dat σ_A^2 en σ_B^2 niet significant verschillen. We gaan dit nu toetsen.

Oplossing

De toetsingsprocedure verloopt als volgt:

1. Bestaat er een verschil in variantie tussen het siliciumgehalte van twee partijen gietijzeren staven?

2. H_0: $\sigma_1^2 = \sigma_2^2$ en H_1: $\sigma_1^2 \neq \sigma_2^2$.

3. $\alpha = 0{,}05$ (tweezijdig).

4. Onder (aanname van) H_0 volgt toetsingsvariabele $F = \frac{S_1^2}{S_2^2}$, een F-verdeling met $\nu_1 = n_1 - 1$ en $\nu_2 = n_2 - 1$ vrijheidsgraden.

5. Berekening van de waarde f van F:

$s_1^2 = 1{,}254$ en $s_2^2 = 0{,}889$, zodat de waarde van de toetsingsvariabele wordt: $f = \frac{1{,}254}{0{,}889} = 1{,}41$, met $\nu_1 = 8 - 1 = 7$ en $\nu_2 = 10 - 1 = 9$ vrijheidsgraden.

6. De overschrijdingskans van F bij $f = 1{,}41$ vinden we met behulp van de F-tabel bij $\nu_1 = 7$ en $\nu_2 = 9$.
We vinden: $P(F > 1{,}41) > 0{,}05$ (eenzijdig) of $P(F > 1{,}41) > 0{,}10$ (tweezijdig).
7. De gevonden overschrijdingskans is duidelijk groter dan $\alpha = 0{,}05$. Conclusie: H_0 wordt niet verworpen.
8. Op grond van het onderzoek kunnen we niet aannemen dat de twee steekproefvarianties significant verschillen.

9.4.4 Het vergelijken van twee fracties

Het komt vaak voor dat twee fracties met elkaar vergeleken moeten worden.

Voorbeeld 16

Een geneesmiddelenfabrikant wenst een onderzoek in te stellen naar de werking van een nieuw medicijn tegen een bepaalde tropische ziekte. Daartoe werd aan een redelijk grote groep personen, die zich ter bestrijding van de bedoelde ziekte door een arts lieten behandelen gevraagd hun medewerking aan dit onderzoek te verlenen. Een groep van 300 patiënten die daarop positief gereageerd hebben, wordt op aselecte wijze onderverdeeld in twee groepen van elke 150 personen, aan te duiden als groep 1 en groep 2. De 150 patiënten van groep 1 krijgen het nieuwe medicijn toegediend, de 150 personen van groep 2 – de zogenaamde *controlegroep* – het traditionele geneesmiddel. Van de 150 patiënten van groep 1 blijken na een van tevoren vastgesteld tijdsverloop 120 personen genezen te zijn. Voor de controlegroep is dit aantal 100. Kan uit deze gegevens geconcludeerd worden dat het nieuwe medicijn effectiever is dan het traditionele medicijn?

Oplossing

Duiden we voor het nieuwe respectievelijk oude medicijn de kans op genezing na een tijdsverloop aan met p_1 respectievelijk p_2 en noemen we K_1 respectievelijk K_2 het aantal genezen patiënten in een steekproef van n_1 respectievelijk n_2 personen, dan zijn K_1 en K_2 binomiaal verdeeld met de parameters n_1 en p_1, respectievelijk n_2 en p_2.
Als n_1 en n_2 voldoende groot zijn (ga dit altijd eerst na), dan kunnen deze binomiale verdelingen benaderd worden door normale verdelingen met als gemiddelde $\mu_1 = n_1 p_1$ respectievelijk $\mu_2 = n_2 p_2$.
Voor de varianties geldt $\sigma_1^2 = n_1 p_1(1 - p_1)$ respectievelijk $\sigma_2^2 = n_2 p_2(1 - p_2)$.
Daar p_1 en p_2 onbekend, zijn gebruiken we de *schatters* voor p_1: $\hat{P}_1 = \dfrac{K_1}{n_1}$, respectievelijk voor p_2: $\hat{P}_2 = \dfrac{K_2}{n_2}$. Ook deze schatters voor p_1 en p_2 zijn normaal verdeeld, met
$\mu_1 = \frac{n_1 p_1}{n_1} = p_1$ en $\sigma_1^2 = \frac{n_1 p_1(1-p_1)}{n_1^2} = \frac{p_1(1-p_1)}{n_1}$, respectievelijk
$\mu_2 = p_2$ en $\sigma_2^2 = \frac{p_1(1-p_1)}{n_1}$.
We willen nu toetsen of er een verschil bestaat tussen de beide fracties. Beter gezegd, of het 'nieuwe' medicijn effectiever is dan het 'oude' medicijn.

Als toetsingsvariabele nemen we het verschil van de fracties in beide steekproeven $\hat{P}_V = \hat{P}_1 - \hat{P}_2$. Mits $\hat{P}_1$ en $\hat{P}_2$ onafhankelijk zijn, kan ook deze nieuwe kansvariabele $\hat{P}_V$ door een normale verdeling worden benaderd met:

$$\mu_V = \mu_1 - \mu_2 = p_1 - p_2 \tag{9.10}$$

$$\sigma_v^2 = \sigma_1^2 + \sigma_2^2 = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2} \tag{9.11}$$

Met behulp van bovenstaande gegevens kan de toetsingsprocedure worden uitgevoerd.

1. Is medicijn 1 effeciever dan medicijn 2?
2. H_0: $p_1 = p_2$ en H_1: $p_1 > p_2$.
3. $\alpha = 0{,}05$ (eenzijdig).
4. De toetsingsvariabele $\hat{P}_V = \hat{P}_1 - \hat{P}_2$ volgt onder H_0 ($p_1 = p_2$) een normale verdeling met $\mu_V = p_1 - p_2 = 0$ en $\sigma_V^2 = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2} = p_1(1-p_1)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$.
5. De hierin voorkomende p_1 is niet bekend. Er zijn twee schattingen voor p_1 te geven: $\hat{P}_1 = \dfrac{k_1}{n_1} = \frac{120}{150} = 0{,}8$ en (onder aanname dat de fractie van de controlegroep dezelfde is als die van de andere groep) $\hat{P}_1 = \dfrac{k_2}{n_2} = \frac{100}{150} = 0{,}667$. Een goede schatting voor P_V kan nu verkregen worden door de beide schattingen gewogen te middelen:

$$\hat{P}_V = \frac{n_1 \cdot \frac{k_1}{n_1} + n_2 \cdot \frac{k_2}{n_2}}{n_1 + n_2} = \frac{k_1 + k_2}{n_1 + n_2} \tag{9.12}$$

Invulling van de gegevens levert $\hat{P}_V = \frac{120+100}{150+150} = 0{,}733$.
De verdeling van de toetsingsvariabele $\hat{P}_V$ is nu bekend. Deze is normaal verdeeld met $\mu_V = 0$ en $\sigma_V = \sqrt{0{,}733(1-0{,}733)\left(\frac{1}{150} + \frac{1}{150}\right)} = 0{,}051$.
6. De waarde van de toetsingsvariabele $\hat{P}_V$ is $0{,}8 - 0{,}667 = 0{,}133$. De rechteroverschrijdingskans onder H_0 bepalen we als volgt:
$P(\hat{P}_V > 0{,}133) = P(U > \frac{0{,}133-0}{0{,}051}) = P(U > 2{,}61) = 0{,}0045$ (eenzijdig).
7. De gevonden overschrijdingskans is kleiner dan $\alpha = 0{,}05$, dus H_0 wordt verworpen ten gunste van. H_1.
8. Het 'nieuwe' medicijn is effectiever dan het 'oude' medicijn.

9.5 De Chi-kwadraattoets voor verdelingen

Met de chi-kwadraattoets (of χ^2-toets) kan men onder andere toetsen of een verdeling van meetuitkomsten een bepaalde theoretische verdeling (bijvoorbeeld normale, binomiale, Poisson- of uniforme verdeling) volgt. De procedure is als volgt.

Stel we beschikken over een (grote) steekproef, waarvan de waarnemingsuitkomsten zijn gegroepeerd in een frequentieverdeling met m klassen. We willen nagaan of de waarnemingsuitkomsten afkomstig zijn uit een bepaalde (theoretische) verdeling. Over de vorm van de verdeling is een veronderstelling te maken en deze veronderstelling stelt men dan ook in de nulhypothese. Op grond van de veronderstelde verdeling worden de verwachte frequenties in een aantal klassen berekend. Vervolgens wordt onderzocht (getoetst) of er een significant verschil bestaat tussen de gevonden frequentie in de verschillende klassen en de verwachte frequentie op grond van de theoretische verdeling.
In alle gevallen wordt als toetsingsvariabele gebruikt:

$$\chi^2 = \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i} \tag{9.13}$$

waarin: m = aantal klassen in de frequentieverdeling.
O_i = waargenomen (*Observed*) frequentie in klasse i.
E_i = verwachte (*Expected*) frequentie in klasse i onder de nulhypothese.
De toetsingsvariabele χ^2 volgt onder de nulhypothese bij benadering een χ^2-verdeling met: $\nu = m - g$ vrijheidsgraden (het bewijs wordt achterwege gelaten). Hierin is g het aantal gegevens dat uit de waargenomen frequentieverdeling berekend moet worden om de theoretische frequentieverdeling te kunnen bepalen.
Voor toepassing van de χ^2-benadering moet aan enkele voorwaarden zijn voldaan:

a. de waarnemingsuitkomsten moeten uit één populatie afkomstig zijn;
b. we moeten ervoor zorgen dat de verwachte (= theoretische) frequentie E_i in iedere klasse groter dan 5 is.

Geldt nu dat voor één of meer klassen dat $E_i \leq 5$, dan kunnen we het beste de verwachte frequentie van de betreffende klasse bij die van de naastliggende klasse(n) optellen en daarmee doorgaan tot voor iedere klasse geldt: $E_i > 5$. Uiteraard moeten ook de corresponderende klassen van de waargenomen frequenties op dezelfde manier worden behandeld. De χ^2-toets passen we toe op de waarnemingsreeks met het verminderde aantal klassen. Is het aantal klassen op bovenvermelde wijze verminderd tot n dan geldt voor het aantal vrijheidsgraden: $\nu = n - g$. Het is hoe dan ook nodig om 1 van n af te trekken, omdat er bij het berekenen van de verwachte frequenties voor gezorgd moet worden dat de som van alle berekende frequenties gelijk is aan de som van alle waargenomen frequenties. Wanneer ook nog eerst a parameters geschat moeten worden om de theoretische frequentieverdeling te kunnen definiëren, gaan nog eens a vrijheidsgraden verloren. Er geldt dus $\nu = n - (a + 1)$.

Opmerking

Wanneer een verdeling op normaliteit getoetst moet worden, moeten eerst μ en σ geschat worden voordat we de verwachte frequenties kunnen berekenen. In totaal gaan er in dat geval dus $2 + 1 = 3$ vrijheidsgraden verloren: $\nu = n - 3$.

Voorbeeld 17

We willen de zuiverheid van een dobbelsteen onderzoeken. De kansverdeling van het aantal ogen van een zuivere dobbelsteen volgt een rechthoekige of uniforme verdeling. (Elk aantal ogen heeft dezelfde kans om op te treden.) We werpen de dobbelsteen 120 keer en noteren bij iedere worp de uitkomst van het aantal ogen. Op grond van de uniforme verdeling verwachten we in elke klasse $\frac{120}{6} = 20$ uitkomsten. De volgende uitkomsten zijn verkregen:

klasse	1	2	3	4	5	6
gevonden freq. (O_i)	16	19	27	17	23	18
verwachte freq. (E_i)	20	20	20	20	20	20

Per klasse (cel) wordt de gevonden frequentie vergeleken met de verwachte frequentie (= 20).

Als de dobbelsteen zuiver is, volgt er: $\chi^2 = \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i}$ is een χ^2-verdeling met $\nu = m - 1$ vrijheidsgraden. Immers, er hoeven geen parameters uit de frequentieverdeling geschat te worden. In de formule $\nu = m - (a + 1)$ is a dus 0.

De toetsingsprocedure in bovengenoemd voorbeeld is als volgt.

1. Is de dobbelsteen zuiver? We toetsen in wezen of de verdeling van de uitkomsten van de dobbelsteen een uniforme verdeling bezit.

2. H_0: dobbelsteen is zuiver en H_1: dobbelsteen is niet zuiver

3. Keuze onbetrouwbaarheid: $\alpha = 0{,}05$ (eenzijdig).
De chi-kwadraattoets is altijd eenzijdig, daar alleen te grote χ^2-waarden leiden tot het verwerpen van de nulhypothese.

4. Uit te voeren toets is de chi-kwadraattoets, met $\chi^2 = \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i}$.

5. Berekening van de waarde c van toetsingsvariabele χ^2:

O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
16	20	-4	16	0,80
19	20	-1	1	0,05
27	20	+7	49	2,45
17	20	-3	9	0,45
23	20	+3	9	0,45
18	20	-2	4	0,20
			$\sum$	4,40

6. In de chi-kwadraattabel wordt de overschrijdingskans van de toetsingsvariabele onder de nulhypothese bij $\nu = 5$ bepaald. In tabel B6 vinden we een overschrijdingskans van: $0{,}10 < P(\chi^2 > 4{,}40) < 0{,}50$.

7. Uit stap 6 concluderen we dat de overschrijdingskans van de toetsingsvariabele groter is dan $\alpha = 0{,}05$. Conclusie: De nulhypothese wordt niet verworpen.

8. Er is geen reden om aan te nemen dat de dobbelsteen niet zuiver is.

9.6 Het toetsen van onafhankelijkheid in een contingentietabel

Contingentietabellen zijn tabellen die worden gebruikt om de onafhankelijkheid te toetsen tussen twee kenmerken.
Van belang bij contingentietabellen zijn de randtotalen, die de marginale verdelingen van de beide kenmerken afzonderlijk weergeven en waarvan de theoretische waarden onder de nulhypothese al of niet bekend zijn.

Voorbeeld 18
We willen onderzoeken of de kleur van het haar en de kleur van de ogen onafhankelijk zijn. We nemen daartoe een steekproef van 600 personen. Iedere persoon zal nu een plaats in een van de rijen krijgen (haarkleur) en een plaats in een van de kolommen (kleur van de ogen). De 'simultane verdeling' van de steekproef is weergegeven in onderstaande tabel.

	haarkleur				
kleur ogen	blond	bruin	zwart	rood	$\sum$
blauw	60	40	60	40	200
grijs	20	50	20	10	100
licht bruin	10	50	10	30	100
bruin	10	160	10	20	200
$\sum$	100	300	100	100	600

We voeren nu de volgende symbolen in:

k: aantal rijen
m: aantal kolommen
n_i: totaal aantal waarnemingen in de i-de rij
n_j: totaal aantal waarnemingen in de j-de kolom
n: totale steekproefgrootte

Wanneer twee kenmerken X_i en Y_j onafhankelijk zijn, geldt er:

$P(X_i \cap Y_j) = P(X_i) \cdot P(Y_j) = \frac{n_i}{n} \cdot \frac{n_j}{n}$.

Anderzijds geldt $P(X_i \cap Y_j) = \frac{n(X_i \cap Y_j)}{n}$, dus geldt $E_{ij} = n \cdot \frac{n_i \cdot n_j}{n^2} = \frac{n_i \cdot n_j}{n}$.

Onder de nulhypothese, dat de twee kenmerken onafhankelijk zijn, vinden we de verwachte frequenties (E_{ij}) in iedere cel dus als volgt: $E_{ij} = \frac{n_i \cdot n_j}{n}$.

Het verwachte aantal in cel (1,1) is bijvoorbeeld: $E_{11} = \frac{200 \times 100}{600} = 33{,}3$

Dit uitgevoerd voor alle cellen, levert de volgende tabel, waarbij het verwachte aantal van iedere cel tussen haakjes is geplaatst.

	haarkleur				
kleur ogen	blond	bruin	zwart	rood	$\sum$
blauw	60 (33,3)	40 (100)	60 (33,3)	40 (33,3)	200
grijs	20 (16,7)	50 (50)	20 (16,7)	10 (16,7)	100
licht bruin	10 (16,7)	50 (50)	10 (16,7)	30 (16,7)	100
bruin	10 (33,3)	160 (100)	10 (33,3)	20 (33,3)	200
$\sum$	100	300	100	100	600

De toetsingsvariabele $\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ volgt onder H_0 bij benadering een chi-kwadraatverdeling met $\nu = (k-1)(m-1)$ vrijheidsgraden.

De toetsingsprocedure voor de toets voor de afhankelijkheid in een contingentietabel gaat nu als volgt.

1. Bestaat er een afhankelijkheid tussen de haarkleur en de kleur van de ogen?
2. H_0: de twee kenmerken zijn onafhankelijk en H_1: de twee kenmerken zijn niet onafhankelijk
3. $\alpha = 0{,}05$ (altijd eenzijdig).
4. De toetsingsvariabele: $\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ volgt onder H_0 een chi-kwadraatverdeli
met: $\nu = (k-1)(m-1)$ vrijheidsgraden.
5. In onderstaande tabel worden alle termen in de somformule berekend:

O_{ij}	E_{ij}	$O_{ij} - E_{ij}$	$\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$
60	33,3	26,7	21,41
40	100	-60,0	36,00
60	33,3	26,7	21,41
40	33,3	6,7	1,35
20	16,7	3,3	0,65
50	50	0	0
20	16,7	3,3	0,65
10	16,7	-6,7	2,69
10	16,7	-6,7	2,69
50	50	0	0
10	16,7	-6,7	2,69
30	16,7	13,3	10,59
10	33,3	-23,3	16,30
160	100	60,0	36,00
10	33,3	-23,3	16,30
20	33,3	-13,3	5,31
600	600	$\sum$	174,04

Conclusie: $c = 174{,}04$ en $k = 4$ en $m = 4$, dus: $\nu = (4-1)(4-1) = 3 \times 3 = 9$ vrijheidsgraden.

6. De overschrijdingskans van $\chi^2 = 174{,}04$ bepalen in de chi-kwadraattabel (B6) bij $\nu = 9$ levert: $P(\chi^2 > 174{,}04) < 0{,}005$.
7. De overschrijdingskans van $\chi^2 = 174{,}04$ is veel kleiner dan $\alpha = 0{,}05$. H_0 wordt daarom verworpen ten gunste van H_1.
8. Er blijkt een afhankelijkheid te bestaan tussen de haarkleur en de kleur van de ogen.

We kunnen nu nog nagaan waar de afhankelijkheden zijn ontstaan. Daartoe kijken we naar de cellen die de hoogste 'chi-kwadraatbijdrage' hebben. De cellen (1,1), (1,2) en (4,2) hebben de hoogste bijdrage. Voor cel (1,1) betekent dit dat er meer personen blauwe ogen hebben met blond haar dan 'verwacht'. Deze analyse kunnen we nu ook maken voor de andere hoge 'chi-kwadraatbijdragen'.

Voorbeeld 19

Drie verschillende materialen worden blootgesteld aan extreme temperaturen. We gaan na of de materialen verkruimelen bij deze blootstellingen. De resultaten zijn weergegeven in de volgende tabel.

	materiaal			
	A	B	C	$\sum$
verkruimeld	41	27	22	90
niet verkruimeld	79	53	78	210
$\sum$	120	80	100	300

De verwachte frequenties zijn:

$E_{11} = \frac{90\times120}{300} = 36 \quad E_{12} = \frac{90\times80}{300} = 24 \quad E_{13} = \frac{90\times100}{300} = 30$

$E_{21} = \frac{210\times120}{300} = 84 \quad E_{22} = \frac{210\times80}{300} = 56 \quad E_{23} = \frac{210\times100}{300} = 70$

Berekening van de toetsingsvariabele:

O_{ij}	E_{ij}	$O_{ij} - E_{ij}$	$\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$
41	36	5	0,69
27	24	3	0,38
22	30	−8	2,13
79	84	−5	0,30
53	56	−3	0,16
78	70	8	0,91
		$\sum$	4,57

χ^2 is Chi-kwadraat-verdeeld met waarde $c = 4{,}57$ en met $\nu = (3-1)(2-1) = 2$ vrijheidsgraden.

Toetsingsprocedure:
- De overschrijdingskans $P(\chi^2 > 4{,}57)$ zoeken we op in tabel B6 bij $\nu = 2$ vrijheidsgraden. Dit levert: $0{,}10 < P(\chi^2 > 4{,}57) < 0{,}25$
- De overschrijdingskans van χ^2 bij een waarde $c = 4{,}57$ is dus groter dan $\alpha = 0{,}05$.
- Conclusie H_0 wordt niet verworpen
- Er is geen reden om aan te nemen, dat de kansen op verkruimeling voor de drie materialen, als ze worden blootgesteld aan extreme temperaturen, verschillend zijn.

9.7 Vergelijking van twee of meer frequentieverdelingen

Tot slot zullen we de chi-kwadraattoets toepassen op het geval, dat we twee of meer steekproeven hebben en we willen toetsen of deze steekproeven afkomstig zijn uit eenzelfde

populatie. Deze toets gebruiken we als de waarnemingsuitkomsten *niet* (bij benadering) normaal verdeeld zijn of in die gevallen waar de verdeling onbekend is.
De toetsingsvariabele is bij deze toets volkomen analoog aan die, welke besproken is in de vorige paragraaf. De interpretatie is echter verschillend en de berekening van de verwachte frequenties verloopt anders.
Stel we hebben k steekproeven met respectievelijk een omvang van $n_1, n_2,.., n_k$ stuks. Voor elk van de k steekproeven zijn de waarnemingsuitkomsten O_{ij} gegroepeerd in m klassen.

steekproef-nummer	klasse 1	...	m	steekproef-grootte
1	O_{11}	...	O_{1m}	n_1
2	O_{21}	...	O_{2m}	n_2
.				.
.		...		.
.				.
k	O_{k1}	...	O_{km}	n_k
$\sum$	O_1	...	O_m	n

We stellen nu de volgende nulhypothese:
H_0: De k steekproeven zijn afkomstig uit dezelfde populatie.
Als alternatieve hypothese geldt:
H_1: De k steekproeven zijn afkomstig uit verschillende populaties.
De kans dat een waarnemingsuitkomst, onder de nulhypothese, in klasse j valt, moet uit de waarnemingsuitkomsten worden geschat. We vinden:

$$p = \frac{1}{n}\sum_{i=1}^{k} O_{ij} = \frac{O_j}{n}$$

zodat de verwachte frequentie van klasse j in steekproef i (= cel (i, j)) wordt geschat door:
$E_{ij} = \dfrac{n_i \cdot O_j}{n}$
waarin:
n_i = rij totaal van de i-rij, dus de grootte van de steekproef i
O_j = kolomtotaal van de j-de kolom.
De toetsingsvariabele wordt weer: $\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{m} \dfrac{(O_{ij} - E_{ij})^2}{E_{ij}}$
waarbij χ^2, onder de nulhypothese, een chi-kwadraatverdeling volgt met $(k-1)(m-1)$ vrijheidsgraden.

Voorbeeld 20

Een bandenfirma wil van 4 verschillende typen banden de duurzaamheid nagaan. Men doet een rijtest en bepaalt na hoeveel kilometers de banden zijn versleten. De uitkomsten zijn:

	aantal kilometers (in duizendtallen)			
type	<30.000	30.000-45.000	>45.000	totaal
A	26	118	56	200
B	23	93	84	200
C	15	116	69	200
D	32	121	47	200
totaal	96	448	256	800

De verwachte frequenties E_{ij} zijn:

$E_{11} = E_{21} = E_{31} = E_{41} = \frac{96 \times 200}{800} = 24$

$E_{12} = E_{22} = E_{32} = E_{42} = \frac{448 \times 200}{800} = 112$

$E_{13} = E_{23} = E_{33} = E_{43} = \frac{256 \cdot 200}{800} = 64$

De toetsingsprocedure verloopt als volgt.

1. Is er verschil in duurzaamheid tussen de bandentypen A, B, C en D?

2. H_0: Er is geen verschil in duurzaamheid tussen de 4 typen A, B, C en D (de steekproeven komen feitelijk uit eenzelfde populatie) en H_1: Er zijn verschillen in duurzaamheid tussen de 4 typen A, B, C en D.

3. $\alpha = 0{,}05$ (eenzijdig).

4. De toetsingsvariabele is de Chi-kwadraat-verdeelde variabele χ^2 met waarde $c = \sum_{i=1}^{k} \sum_{j=1}^{m} \dfrac{(O_{ij} - E_{ij})^2}{E_{ij}}$.

Berekening van c gaat als volgt (zie de volgende tabel).

5. De overschrijdingskans van de toetsingsvariabele χ^2 bij een waarde $c = 22{,}79$ opzoeken in tabel B6 bij $\nu = (k-1)(m-1) = (3-1)(4-1) = 6$ vrijheidsgraden levert $P(\chi^2 > 22{,}79) < 0{,}005$.

6. De overschrijdingskans van χ^2 is kleiner dan $\alpha = 0{,}05$. Conclusie: H_0 wordt verworpen ten gunste van H_1

7. Op grond van de gehouden steekproef kunnen we concluderen dat er een verschil in duurzaamheid is tussen de 4 bandentypen. De verschillen vinden we door te kijken naar de cellen met de hoogste chi-kwadraatbijdragen. Zo heeft cel (2,3) een significante afwijking. De frequentie is hoger dan wat men theoretisch kan verwachten. De afwijking is positief. Conclusie is dat de banden van type B langer meegaan. Ook cel (4,3) heeft een grote chi-kwadraatbijdrage, maar de afwijking is negatief. De banden van type D gaan dus korter mee.

cel	O_{ij}	E_{ij}	$O_{ij} - E_{ij}$	$\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$
11	26	24	2	0,17
12	118	112	6	0,32
13	56	64	−8	1,00
21	23	24	−1	0,01
22	93	112	−19	3,22
23	84	64	20	6,25
31	15	24	−9	3,38
32	116	112	4	0,14
33	69	64	5	0,39
41	32	24	8	2,67
42	121	112	9	0,72
43	47	64	−17	4,52
$\sum$	800	800		22,79

9.8 Het toetsen van uitschieters

We besluiten met twee toetsen waarbij we de tot dusver gebruikte kansverdelingen niet meer kunnen gebruiken.

9.8.1 Het verwerken en toetsen van verdachte uitkomsten

Het komt nogal eens voor dat men in een reeks meetuitkomsten één of meer waarden aantreft, die veel afwijken van de overige waarden. De vraag wordt dan gesteld wat met deze verdachte waarden moet of mag worden gedaan. De verdachte waarde weglaten, handhaven of vervangen door een nieuwe meetuitkomst?

Veel onderzoekers hebben de slechte gewoonte om verdachte waarden zonder meer weg te laten of te vervangen. Dit 'zonder-meer-weglaten' is alleen dan toegestaan als men een bewijs in handen heeft dat de afwijkende waarde het gevolg is van een technische fout.

We spreken van een technische fout als de afwijking is ontstaan door oorzaken die niets met het onderzochte materiaal te maken hebben, zoals:

- fouten bij behandeling of bewaring van de monsters;
- fouten bij het voorbereiden van de metingen;
- technische afwijkingen tijdens de metingen;
- fouten bij de verwerking van de meetuitkomsten, afleesfouten, reken- of typefouten, enzovoorts.

Als bewijs voor een technische fout geldt bijvoorbeeld dat men gezien moet hebben dat er tijdens de meting iets misging, of dat aangetoond kan worden dat er een rekenfout is gemaakt, dat een punt van een kromme verkeerd is afgelezen, of dat er 'onmogelijke' uitkomsten zijn verkregen.
Vermoedens, achteraf door onderlinge vergelijking van meetuitkomsten, dat 'er ergens wel wat fout zal zijn gegaan', hebben geen enkele bewijskracht. Heeft men geen bewijs van een technische fout, dan kan men door toetsing nagaan of de verdachte uitkomst statistisch gezien een uitschieter of *uitbijter* is of niet.
Bij verdachte waarden kunnen we de volgende situaties onderscheiden.

a. *Technische fouten*
Bij gemaakte rekenfouten of aflezingen van diagrammen, is het meestal wel mogelijk de gemaakte fouten te corrigeren. Als er geen correctie mogelijk is, worden de foute waarden weggelaten. Als hierdoor het aantal overblijvende uitkomsten te gering wordt, dan kunnen we de geschrapte waarden vervangen door uitkomsten van nieuwe metingen.
Het toetsen of een afwijkende waarde een uitschieter is, komt neer op het berekenen van de overschrijdingskans P voor deze waarde.

overschrijdingskans	conclusie
$P \leq 1\%$	storende uitschieter
$1 < P \leq 5\%$	uitschieter
$P > 5\%$	geen uitschieter

b. *Storende uitschieters ($P \leq 1\%$)*
De storende uitschieters worden bij berekening van gemiddelden, spreidingen, enzovoorts niet meegerekend. Echter bij elke publicatie (onderzoekbriefjes, proefverslagen, enzovoorts) moet uitdrukkelijk worden vermeld, dat men één of meer uitschieters bij de berekeningen heeft weggelaten (ook de waarden vermelden).

c. *Uitschieters ($1 < P \leq 5\%$)*
Deze uitschieters worden wel in de berekeningen meegenomen. Bij elke vorm van publicatie moet uitdrukkelijk worden vermeld dat uitschieters (ter grootte van ...) zijn meegerekend.

d. *Geen uitschieters ($P > 5\%$)*
Is een verdachte uitkomst geen uitschieter – dus $P > 5\%$ – dan wordt de waarde gewoon in alle berekeningen opgenomen, zonder speciale vermelding.

e. *Aanvullende metingen*
Afwijkende waarden met een overschrijdingskans $1 < P \leq 5\%$ mogen dus niet bij de berekeningen worden weggelaten. Vooral in kleine steekproeven kunnen deze afwijkende waarden een sterke invloed uitoefenen op het gemiddelde en de standaardafwijking. Om deze invloed wat te verminderen, is het raadzaam een reeks aanvullende metingen uit te voeren. De oorspronkelijke en de aanvullende meetuitkomsten worden

dan samengevoegd voor verdere verwerking. Indien gewenst, kan men na samenvoeging opnieuw de uitschieters toetsen. Het kan zijn dat de verdachte uitkomst nu wel een overschrijdingskans $P \leq 1\%$ heeft en dus in verdere berekeningen mag worden weggelaten.

9.8.2 De toets van Grubbs

Met deze toets, ontwikkeld door Frank Grubbs, kunnen we nagaan of verdacht hoge of verdacht lage waarden in een reeks meetuitkomsten of in een reeks gemiddelden, statistisch gezien, echte uitschieters zijn.

Procedure

Bij de toets van Grubbs wordt verondersteld dat de meetwaarden afkomstig zijn uit een normaal verdeelde populatie. De meetuitkomsten worden gerangschikt in volgorde van grootte, waarna de waarde van de toetsingsvariabele T (niet te verwarren met T uit de t-verdeling!) als volgt wordt bepaald:

a. voor een verdacht grote waarde:

$$t = \frac{x_{(n)} - \bar{x}}{s}$$

met:

$x_{(n)}$ = hoogste meetwaarde

$\bar{x}$ = gemiddelde van alle meetwaarden

s = standaardafwijking van alle meetwaarden

b. voor een verdacht kleine waarde:

$$t = \frac{\bar{x} - x_{(1)}}{s}$$

met:

$x_{(1)}$ = laagste meetwaarde

In beide gevallen wordt de overschrijdingskans van de toetsingsvariabele opgezocht in een tabel, afhankelijk van het aantal meetwaarden n.

Voorbeeld 21

Bij een onderzoek zijn de volgende 6 uitkomsten verkregen: 127 118 125 127 127 128. De onderzoeker vindt de waarde 118 verdacht klein en wil dit toetsen.

Oplossing

De 6 meetwaarden worden gerangschikt in volgorde van grootte en vervolgens wordt het gemiddelde en de standaardafwijking bepaald.

118	125	126	127	127	128
$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$

$\bar{x} = 125{,}1667$ en $s_X = 3{,}656$

De toetsingsvariabele heeft een waarde:

$$t = \frac{\bar{x} - x_{(1)}}{s_X} = \frac{125{,}1667 - 118}{3{,}656} = 1{,}96$$

In de tabel voor de toets van Grubbs (tabel B10) wordt de overschrijdingskans van T bij een waarde $t = 1{,}96$ opgezocht bij $n = 6$.
Dit levert: $P(T > 1{,}96) < 0{,}01$
Conclusie: de uitkomst 118 is een storende uitschieter (niet meerekenen, wel noemen).

Voorbeeld 22

Van een serie van 10 meetwaarden zijn de volgende uitkomsten verkregen:

10,3	10,5	10,6	10,6	10,9	10,9	11,3	11,5	11,8	13,2

Is 13,2 een uitschieter?

Oplossing

We berekenen eerst $\bar{x}$ en s van de 10 uitkomsten: $\bar{x} = 11{,}15$ en $s = 0{,}859$.
Omdat 13,2 een 'verdacht' hoge waarde is, berekenen we de waarde van de toetsingsvariabele T als volgt:

$$t = \frac{x_{(n)} - \bar{x}}{s} = \frac{13{,}2 - 2{,}39}{0{,}859} = 2{,}39.$$

In tabel B8 vinden we bij $n = 6$ de overschrijdingskans van T bij $t = 2{,}39$.
Dit levert: $0{,}01 < P(T > 3{,}39) < 0{,}025$.
Conclusie: de waarde 13,2 is een uitschieter, maar geen storende uitschieter (wel noemen, niet meerekenen).

9.8.3 De toets van Cochran (voor verdacht grote varianties)

Met deze toets kunnen we nagaan of een verdacht grote variantie ($s^2_{\max}$) in een groep van k varianties een echte uitschieter is of niet.
Voorwaarde voor het gebruik van deze toets is dat de k steekproeven, waarvan de varianties berekend worden, even groot zijn en afkomstig zijn uit (bij benadering) normaal verdeelde populaties. De toetsingsvariabele voor de toets van Cochran is T met waarde:

$$t = \frac{s^2_{\max}}{\sum_{i=1}^{k} s_i^2}$$

k = aantal varianties, inclusief de verdachte variantie
ν = aantal vrijheidsgraden van elk der k varianties.
De toets demonstreren we aan de hand van het volgende voorbeeld.

Voorbeeld 23

In een onderzoek zijn 5 spoelen garen betrokken. Aan elke spoel worden 10 metingen gedaan. Behalve het gemiddelde van elke spoel wordt ook de variantie berekend. De vijf varianties zijn: $s_1^2 = 26 \quad s_2^2 = 40 \quad s_3^2 = 83 \quad s_4^2 = 24 \quad s_5^2 = 28.$

De variantie van spoel 3 ($s_3^2 = 83$) vindt men verdacht groot in vergelijking met de andere vier varianties. De verdacht grote variantie wordt onderzocht met behulp van de toets van Cochran.

Voor de toetsingsvariabele T vinden we een waarde:

$$T = \frac{s_{\max}^2}{\sum_{i=1}^{k} s_i^2} = \frac{83}{26 + 40 + 83 + 24 + 28} = 0{,}413 \text{ met: } k = 5 \text{ en } \nu = 10 - 1 = 9.$$

De overschrijdingskans van $T = 0{,}413$, wordt bepaald met behulp van de kritieke waarden voor T uit tabel B11. Voor $k = 5$ en $\nu = 9$ vinden we:

$\alpha = 5\%$: $k_{0,05} = 0{,}424$.

$\alpha = 1\%$: $k_{0,01} = 0{,}485$.

Vergelijken we de waarde van toetsingsvariabele T ($= 0{,}413$) met de beide kritieke waarden $k_{0,05} = 0{,}424$ en $k_{0,01} = 0{,}485$, dan blijkt dat de overschrijdingskans bij $t = 0{,}413$ groter is dan 5%.

Hieruit moeten we concluderen dat de verdachte variantie $s_3^2 = 83$ net geen uitschieter is.

Opmerking

Voor gelijkblijvende waarden van de overige 4 varianties, zal s_3^2 pas een significante uitschieter zijn als $s^2 > 87$. Immers, bij een overschrijdingskans van 5% behoort een kritieke waarde van 0,424. Hierbij kan nu de kritieke waarde van de toetsingsvariabele (=87) worden berekend.

Opgaven

1. Een kledingmagazijn verkoopt kostuums van maat 51, waarvan de fabrikant als norm heeft opgegeven voor de lengte van de broek 112 cm, met $\sigma = 2$ cm.

Men bemerkt echter dat de broeken vaak te kort zijn, wat tot het vermoeden leidt dat de lengte van de broeken systematisch te kort is.

In het kledingmagazijn besluit men tien pantalons van maat 51 willekeurig aan een partij te onttrekken en na te meten. Men krijgt de volgende gegevens (in cm):

110	108	113	112	109	107	108	112	113	110

Zijn de broeken van maat 51 tekort (toets met $\alpha = 0{,}05$)?

2. Volgens de fabrikant van een nieuw type personenauto verbruikt deze auto bij een constante snelheid van 90 km/uur gemiddeld 7,1 liter benzine per 100 gereden kilometers. Door een consumentenorganisatie werd van 10 van zulke auto's het benzineverbruik gemeten. Men vond (in liters per 100 km):

7,3	7,0	7,3	7,7	7,4	7,2	7,3	6,9	7,4	7,5

Toets met een onbetrouwbaarheid 0,05, de hypothese dat de bewering van de fabrikant juist is, onder de aanname dat het benzineverbruik in liters per 100 km normaal verdeeld is.

3. Men wil twee merken kooktoestellen met elkaar vergelijken. De kwaliteit van een kooktoestel wordt onder meer bepaald door het warmterendement, dat wil zeggen de verhouding tussen de benutte en de vrijgekomen warmte. Men heeft van de merken A en B een aantal toestellen getest en de volgende rendementen (in %) gevonden. Van merk A hebben we slechts 5 metingen doordat 1 meting is uitgevallen.

A:	59	55	61	58	60	
B:	63	64	59	60	65	60

Bestaan er verschillen in rendement tussen de merken A en B (toets met $\alpha = 0{,}05$)?

4. Tijdens een chemisch proces wordt een bepaalde hoeveelheid grondstof omgezet in een hoeveelheid eindproduct. Bij een volledige chemische omzetting spreken we van een chemisch rendement van 100%. In de praktijk wordt dat evenwel nooit bereikt, doch men wil het zo hoog mogelijk opvoeren. Er zijn nu twee apparaten met verschillende constructies. Men wil onderzoeken of dit verschil in constructie invloed heeft op het rendement. Men maakt gedurende acht dagen, per dag een charge grondstof aan en verdeelt die in twee porties. Eén portie wordt op apparaat A gedraaid en de andere portie op apparaat B, vervolgens wordt het rendement voor beide apparaten bepaald.

dag	A	B
1	89,3	92,6
2	87,5	90,3
3	91,4	91,2
4	88,1	92,6
5	88,2	85,8
6	91,7	95,8
7	83,7	82,6
8	87,3	91,6

Bestaat er een verschil in rendement tussen de beide apparaten? (toets met $\alpha = 0{,}05$)

5. Een meetmethode wordt door twee analisten uitgevoerd. Beide analisten verrichten 10 metingen aan eenzelfde standaardmonster. Van beide meetseries wordt de standaardafwijking bepaald:
$s_A = 1{,}44$ mm en $s_B = 2{,}87$ mm
Bestaat er een verschil in meetnauwkeurigheid tussen de beide analisten? (toets met $\alpha = 0{,}05$)

6. Bij de bepaling van de viscositeit van een hars zijn de volgende waarden verkregen:

1,20	1,28	1,30	1,45	1,25	1,23	1,29	1,30	1,28	1,17	1,10

Ga na of bij bovenstaande uitkomsten storende uitschieters zitten.

7. Van 6 steekproeven die ieder bestaan uit 4 waarnemingen, heeft men de variantie bepaald.

De varianties zijn:	22,8	25,0	30,4	27,9	79,2	33,1

Ga na of 79,2 een storende uitschieter is.

8. Men beschikt over de volgende steekproefresultaten:

steekproeven				
A	B	C	D	E
2,5	3,5	2,7	3,8	5,7
3,0	3,1	3,1	2,2	5,8
2,9	3,7	2,3	3,5	4,9
3,0	3,5	2,9	3,9	2,9
2,8	2,3	3,1	3,1	5,9

Bereken de variantie van de steekproeven en toets of er een te grote variantie bij is. Zo ja, waardoor kan die zijn ontstaan?

9. Een product wordt geproduceerd in een reactieketel, door drie grondstoffen met elkaar te laten reageren, onder invloed van een bepaalde hoeveelheid van katalysator. Een belangrijk criterium voor het eindproduct is het gehalte aan een bepaalde stof. Uit onderzoek bleek dat het gehalte aan desbetreffende stof in het eindproduct afhankelijk kan zijn van de hoeveelheid katalysator. Men doet nu een proef om te zien of een verhoging van 3 naar 6 gram katalysator, ook een verhoging van de bepaalde stof in het eindproduct geeft. Bij beide hoeveelheden is $\sigma = 2$. Men toetst daarbij de volgende hypothesen:
Nulhypothese $\mu = 3{,}5$ tegen de alternatieve hypothese $\mu = 6{,}75$ procent van de stof in het eindproduct.
De te gebruiken toetsingsvariabele is het gemiddelde van een steekproef van vier waarnemingen.

a. Veronderstel dat H_0 waar is, specificeer de kansverdeling van de toetsingsvariabele T.
b. Bepaal, onder de voorwaarde dat H_0 waar is, de kans dat $T > 5{,}46$.
c. Veronderstel dat H_1 waar is, specificeer dan de verdeling van T.
d. Bepaal in geval c de kans dat $T > 5{,}46$.
De nulhypothese wordt verworpen bij $T > 5{,}46$.
e. Hoe wordt de waarde $T = 5{,}46$ genoemd?
f. Hoe wordt het gebied $T > 5{,}46$ genoemd, indien H_0 waar is?
g. Hoe groot is de onbetrouwbaarheid van de toets?
h. Hoe groot is het onderscheidingsvermogen van de toets?

10. Een firma in lijmsoorten wil het effect van een reclamecampagne nagaan ten aanzien van de naamsbekendheid van zijn lijm 'Beverkracht' in vergelijking met het concurrentieproduct 'Bisonkracht'. De firma laat een onderzoek verrichten door een onderzoeksbureau, twee weken voor de campagne en twee weken na de campagne.

merk	voor campagne	na campagne
Beverkracht	335	486
Bisonkracht	565	515

Toets of er een effect bestaat door de reclamecampagne ten aanzien van de naamsbekendheid. (Toets met $\alpha = 0{,}05$.)

11. Bij een proef omtrent het afdichten van plastic bakjes worden 3 verschillende afdichtingmethoden A, B en C met elkaar vergeleken. De afdichting wordt gecontroleerd door gevulde en gesloten bakjes, na sterilisatie, gedurende een zekere tijd op te slaan in met pathogene bacteriën besmet water. Daarna wordt de inhoud van elk bakje gecontroleerd op de aanwezigheid van deze bacteriën. De uitslag van het onderzoek is als volgt:

methode	steekpr. grootte	percentage besmette bakjes
A	75	13,3%
B	90	15,5%
C	195	11,3%

Bestaat er een verschil tussen de drie afdichtingmethoden? (Toets met $\alpha = 0{,}05$.)

12. Een fabrikant van computerchips garandeert dat deze bij normaal gebruik een gemiddelde levensduur hebben van meer dan 8000 bedrijfsuren. Van 20 van zulke chips is de levensduur bepaald.
Voor het gemiddelde vond men 8300 uur en voor de standaardafwijking 1000 uur.

Wanneer we veronderstellen dat de levensduur van chips normaal verdeeld is, kunnen we dan de conclusie trekken dat de garantie van de fabrikant juist is? (toets met $\alpha =$ 0,05)

13. In een enquête onder het personeel van een groot bedrijf werd in vraag 17 verzocht een oordeel te geven over de werkomstandigheden op de afdeling waar zij werkzaam zijn. Van de in totaal 539 medewerkers van de afdelingen X, Y en Z hebben er precies 500 aan de enquête meegedaan. Hun reactie op vraag 17 kan als volgt worden samengevat:

	oordeel werkomstandigheden		
afdeling	ontevreden	matig	tevreden
X	36	56	121
Y	32	44	61
Z	48	55	47

Onderzoek of de mate van tevredenheid over de werkomstandigheden afhankelijk is van de afdeling waarop men werkt. Kies een onbetrouwbaarheid $\alpha = 0{,}05$.

14. Volgens de Kwaliteitsdienst van de N.V. Tarwex mag het gemiddelde nettogewicht van een pak tarwevlokken niet minder dan 500 gram bedragen. Aangenomen mag worden dat het bedoelde gewicht bij benadering normaal verdeeld is met een standaardafwijking $\sigma = 28$ gram. Bij een controle leverde een steekproef van 16 pakken een gemiddeld gewicht van 485 gram op. Kan er behoudens een onbetrouwbaarheid $\alpha = 0{,}05$ geconcludeerd worden dat de machine waarop de pakken tarwevlokken worden gevuld, moet worden bijgesteld?

15. In een staalfabriek kan volgens twee methoden, methode A en methode B, betonstaal gewalst worden. Uit ervaring weet men dat de treksterkte van betonstaal normaal verdeeld is, voor methode A met een standaardafwijking van 1200 N en voor methode B met een standaardafwijking van 1600 N. De gemiddelde treksterkte was voor een steekproef van 12 stukken betonstaal, gewalst volgens methode A, gelijk aan 60000 N en voor een steekproef van 15 stukken betonstaal, gewalst volgens methode B, gelijk aan 59000N.
Volgt hieruit dat betonstaal, gewalst volgens methode A niet dezelfde gemiddelde treksterkte heeft als betonstaal, gewalst volgens methode B? Kies $\alpha = 0{,}05$.

16. Een bepaald kwantitatief kenmerk van een bepaald product heeft een standaardafwijking van 20. Na een inmiddels opgeheven storing in het productieproces bleek de bedoelde eigenschap in een steekproef van 16 stuks een standaardafwijking 24 te hebben.
Moet hieruit met onbetrouwbaarheid 0,05 geconcludeerd worden dat de processpreiding na de storing groter is geworden?

17. In een supermarkt krijgen de klanten de gelegenheid om twee verschillende merken soep te proeven. Van een groep van 100 klanten vonden er 63 de soep van merk A lekkerder dan de soep van merk B. Toets met $\alpha = 0{,}10$ de hypothese H_0 dat geen voorkeur voor een van de twee merken soep bestaat tegen de alternatieve hypothese H_1 dat de voorkeur voor merk A groter is dan die voor merk B.

10 Lineaire regressie en correlatierekening

10.1 Inleiding

Regressie-analyse is een methode waarmee we kunnen onderzoeken of er tussen twee (of meer) kansvariabelen een bepaald verband bestaat. Bestaat er een verband tussen het lichaamsgewicht en de lichaamslengte van een bepaalde groep mensen? Is er een relatie tussen de prijs van en de vraag naar een zeker product? Is de opbrengst van een bepaalde akker afhankelijk van de gebruikte hoeveelheid kunstmest en/of de zuurgraad van de grond? Is de druk afhankelijk van het volume en/of de temperatuur? Is de stroomsterkte in een bepaald netwerk gerelateerd aan de ingestelde spanning? Op dit soort vragen kan een antwoord worden gevonden via de methode van de *regressie-analyse*. Met deze methode kunnen we nagaan *óf* er een verband bestaat tussen twee (of meer) kansvariabelen. Want ook kan – indien er inderdaad een verband blijkt te bestaan – worden vastgesteld op welke wijze dit verband in een formule kan worden vastgelegd. Dit is minder eenvoudig dan op het eerste gezicht lijkt. In het algemeen zullen de meetpunten die bij de analyse gebruikt worden niet allemaal op de grafiek van de te zoeken functie liggen. Dit komt niet alleen door meetfouten maar ook door het toevalskarakter van de meetpunten. Er is dan ook geen eenduidig antwoord te geven op de vraag welke functie het verband tussen de variabelen exact weergeeft. Op basis van een bepaald criterium kan wel een functie gevonden worden die het verband tussen de variabelen het *beste* weergeeft. Het meest gebruikte criterium is het zogenaamde *kleinste-kwadraten-criterium*. In dit hoofdstuk zullen we in eerste instantie bekijken hoe de hiermee de formule kan worden gevonden die een *lineair* verband tussen twee variabelen X en Y weergeeft. We zullen echter ook ingaan op de vraag hoe met het kleinste-kwadraten-criterium een benadering kan worden gevonden voor een niet-lineair verband. Deze vorm van regressie-analyse noemt men vaak *curve-fitting*.
Naast de regressie-analyse kennen we ook de *correlatierekening*. Hiermee kan de *mate* van afhankelijkheid tussen twee (of meer) variabelen worden vastgesteld. Zo blijken bijvoorbeeld de eindexamencijfers voor de vakken wiskunde en natuurkunde vaak sterk gecorreleerd te zijn, maar de cijfers voor wiskunde en engels veel minder. Voor zover het de mate

van correlatie tussen twee kansvariabelen met een *lineaire* regressie betreft, worden in dit hoofdstuk de begrippen *correlatiecoëfficiënt* en *covariantie* ingevoerd.

10.2 De methode van de kleinste kwadraten

Stel dat we willen onderzoeken of er een bepaald verband bestaat tussen de variabelen X en Y die we beide kunnen meten aan elk van de n elementen van een bepaalde verzameling (denk hierbij bijvoorbeeld aan de meting van lengte en gewicht van een groep volwassenen). Om een eerste indruk te krijgen of er al dan niet een verband bestaat, kunnen we – zie figuur 10.1 – de n meetpunten (x_i, y_i) $(i = 1, 2, 3, ..., n)$ tegen elkaar uitzetten in een *punten- of scatterdiagram.*

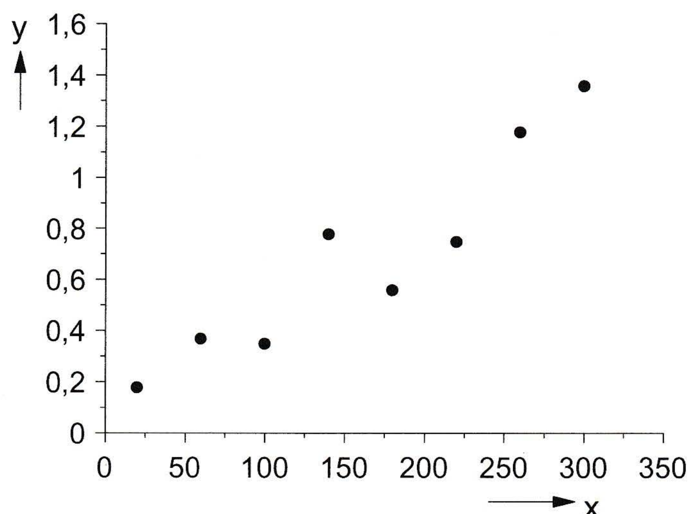


Fig. 10.1 Puntendiagram van een reeks van 8 meetpunten

Op het eerste gezicht blijkt voor de 8 meetpunten van figuur 10.1 het verband tussen X en Y min of meer lineair te zijn. Om hierin meer inzicht te krijgen, zouden we door de puntenwolk een rechte lijn kunnen trekken op een zodanige wijze dat elk van de 8 punten er zo dicht mogelijk bij ligt. De vraag is echter op welke wijze dit gerealiseerd kan worden. Wanneer we de vergelijking van de bedoelde rechte lijn formuleren als:

$$\hat{y} = p(x) = ax + b \qquad (10.1)$$

($\hat{y}$ is het symbool dat hier gebruikt wordt voor de *benadering* van y), blijkt het mogelijk een waarde voor de *richtingscoëfficiënt* a en het *intercept* b te berekenen met behulp van de *methode van de kleinste kwadraten.* Om duidelijk te maken wat deze methode inhoudt, verwijzen we naar fig. 10.2.

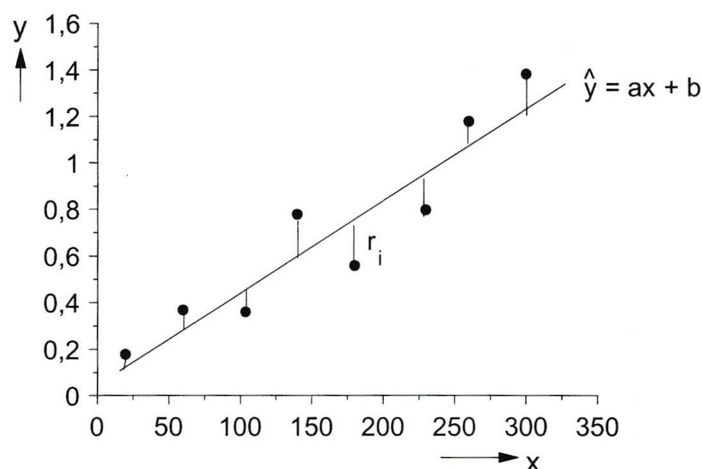


Fig. 10.2 De residuen r_i van de 8 meetpunten van figuur 10.1

Voor elk punt (x_i, y_i) uit de puntenwolk van figuur 10.1 definiëren we het *residu* r_i (ook wel fout of afwijking genoemd) als het verschil tussen de *gemeten* waarde y_i en de door toepassing van de vergelijking $\hat{y} = p(x) = ax+b$ verkregen waarde $\hat{y}_i = p(x_i) = ax_i + b$. Er geldt dan:

$$r_i = y_i - \hat{y}_i = y_i - p(x_i) = y_i - ax_i - b \tag{10.2}$$

Opmerking

Voor elk punt (x_i, y_i) is het residu r_i te beschouwen als de fout in de gemeten waarde y_i van Y ten opzichte van de lijn $\hat{y} = p(x) = ax+b$. Hierbij wordt verondersteld dat x_i een onafhankelijk gekozen of *ingestelde* waarde van X vertegenwoordigt. Daarom noemen we Y de (van X) *afhankelijke* variabele en X de (van Y) *onafhankelijke* variabele.

We zullen veronderstellen dat de r_i voor elke i normaal verdeeld zijn met gemiddelde 0 (en met standaardafwijking σ_r, waarover later meer). Met $r_i = y_i - p(x_i)$ geldt dat $E(r_i) = E(y_i) - E(p(x_i)) = E(y_i) - p(x_i) = 0$. We kunnen dus stellen dat $E(y_i) = p(x_i)$. De waarde $p(x_i) = ax_i + b$ is te beschouwen als de 'meest waarschijnlijke' meetwaarde van Y bij de instelwaarde $X = x_i$.

De rechte lijn $\hat{y} = ax + b$, die de eigenschap bezit dat de som van de kwadraten van alle residuen r_i zo klein mogelijk is, noemen we de *regressielijn* van Y op X.

De methode van de kleinste kwadraten eist dus dat we a en b zodanig kiezen, dat de som van de kwadraten van de n residuen r_i $(i = 1, 2, 3, ..., n)$ minimaal is.

Opmerking

De methode van de kleinste kwadraten is niet de enige methode om een regressielijn te bepalen. We kunnen aan de regressielijn bijvoorbeeld ook de eis stellen dat de som van de absolute waarden van de n residuen zo klein mogelijk is of dat de grootste van de n residuen (in absolute waarde) zo klein mogelijk is. Deze methoden hebben echter zowel praktische als wiskundige bezwaren. Daarom zullen we er in het kader van dit boek niet verder op ingaan.

Om $f(a,b) = \sum_{i=1}^{n} r_i^2$ te kunnen minimaliseren, moeten we de partiële afgeleiden $\frac{\partial f}{\partial a}$ en $\frac{\partial f}{\partial b}$ beide gelijk aan 0 stellen. Doen we dit, dan volgt hieruit een stelsel van twee vergelijkingen en twee onbekenden:

$$a\sum_{i=1}^{n}(x_i) + bn = \sum_{i=1}^{n} y_i \tag{10.3}$$

$$a\sum_{i=1}^{n}(x_i)^2 + b\sum_{i=1}^{n}(x_i) = \sum_{i=1}^{n}(x_i y_i) \tag{10.4}$$

Met $\bar{x}$ als het gemiddelde van alle x-coördinaten en $\bar{y}$ als gemiddelde van alle y-coördinaten is de oplossing van dit stelsel:

$$b = \bar{y} - a\bar{x} \tag{10.5}$$

en:

$$a = \frac{\sum_{i=1}^{n}(x_i y_i) - n(\bar{x})(\bar{y})}{\sum_{i=1}^{n}(x_i)^2 - n(\bar{x})^2} \tag{10.6}$$

Opmerking

Voor de laatste formule kan ook geschreven worden

$$a = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{10.7}$$

Voorbeeld 1

Voor een groep van 6 personen zijn de volgende lichaamslengten en lichaamsgewichten gegeven:

persoon (i)	1	2	3	4	5	6
lengte in cm (x_i)	164	174	176	180	184	190
gewicht in kg (y_i)	54	60	65	71	76	82

We bepalen met behulp van de methode van de kleinste kwadraten de regressiecoëfficiënt a en het intercept b voor de lineaire regressie van Y op X.
Ten behoeve van de berekening van de regressiecoëfficiënt met behulp van formule (10.4) maken we eerst de volgende berekening:

nummer	x_i	y_i	x_i^2	$x_i y_i$
1	164	54	26896	8856
2	174	60	30276	10440
3	176	65	30976	11440
4	180	71	32400	12780
5	184	76	33856	13984
6	190	82	36100	15580
totaal	1068	408	190504	73080

Met $n = 6$ vinden we: $\bar{x} = \dfrac{1068}{6} = 178$ en $\bar{y} = \dfrac{408}{6} = 68$.

Formule (10.6) levert dan op: $a = \dfrac{73080 - 6(178)(68)}{190504 - 6(178)^2} = \dfrac{456}{400} = 1{,}14$.

Met $a = 1{,}14$, $\bar{x} = 178$ en $\bar{y} = 68$ vinden we ten slotte met behulp van formule (10.5): $b = 68 - 1{,}14(178) = -134{,}92$.
De gevraagde regressievergelijking luidt dus: $\hat{y} = 1{,}14x - 134{,}92$.

Opmerking

De in voorbeeld 1 berekende regressielijn is slechts gebaseerd op een beperkt aantal meetpunten (6) binnen een beperkt gebied van de instelvariabele X (164 t/m 190 cm). Daarom zou het onjuist zijn de regressielijn te gebruiken om voor elke willekeurige persoon het lichaamsgewicht te berekenen op basis van een gegeven lichaamslengte.

Opdracht

Controleer de juistheid van de bovenstaande opmerking voor de eigen situatie alsmede voor een kind met een lichaamslengte van 100 cm.

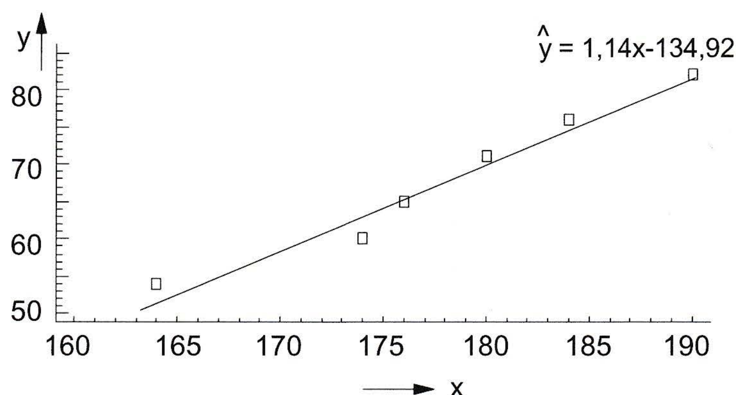


Fig. 10.3 Puntenwolk en regressielijn bij voorbeeld 1

In feite kan men de in voorbeeld 1 berekende regressielijn $\hat{y} = 1{,}14x - 134{,}92$ slechts gebruiken voor de berekening van een schatting van het gemiddelde lichaamsgewicht van personen met een lichaamslengte tussen 164 en 190 cm. En dan nog met de grootst mogelijke voorzichtigheid, want uiteindelijk is deze schatting op slechts 6 waarnemingsparen gebaseerd.

10.3 De tweede regressielijn

In voorbeeld 1 hebben we voor de instelvariabele X de lichaamslengte en voor de meetwaarde Y het lichaamsgewicht van 6 personen gekozen. De aldus berekende regressielijn $\hat{y} = 1{,}14x - 134{,}92$ kan, met de nodige voorzichtigheid, gebruikt worden om voor alle personen met een lichaamslengte x_i tussen 164 en 190 cm (het instelgebied van de variabele X) een schatting te berekenen van het gemiddelde lichaamsgewicht μ_Y. Wensen we het omgekeerde te doen, dus wensen we voor alle personen met een lichaamsgewicht y_i tussen 54 en 82 kg (het meetgebied van de variabele Y) een schatting $\hat{x}$ te berekenen van de gemiddelde lichaamslengte μ_X dan mag hiervoor de regressielijn $\hat{y} = 1{,}14x - 134{,}92$ niet gebruikt worden. In dat geval moeten we de zogenaamde *tweede regressielijn* gebruiken, die van de vorm $\hat{x} = q(y) = cy + d$ is en die dus voor de meetwaarden x_i van X (de lichaamslengte) vastlegt hoe deze afhangen van de instelwaarden y_i van Y (het lichaamsgewicht).
Met rolverwisseling van X en Y kan de vergelijking van de tweede regressielijn uit de formules (10.5) en (10.6) bepaald worden. Wanneer we deze vergelijking aanduiden met $\hat{x} = q(y) = cy + d$, ontstaan de formules:

$$d = \bar{x} - c\bar{y} \qquad (10.8)$$

en

$$c = \frac{\sum_{i=1}^{n}(y_i x_i) - n(\bar{y})(\bar{x})}{\sum_{i=1}^{n}(y_i)^2 - n(\bar{y})^2} \tag{10.9}$$

Voorbeeld 2

Bepaal de vergelijking van de tweede regressielijn voor de 6 meetpunten, waarvan in voorbeeld 1 de vergelijking van de eerste regressielijn is bepaald.

Oplossing

Met $\sum_{i=1}^{n}(y_i x_i) = 73080$, $\bar{x} = 68$ (zie voorbeeld 1) en met

$\sum_{i=1}^{n}(y_i)^2 = 542 + 602 + 652 + 712 + 762 + 822 = 28282$

vinden we volgens formule (10.9):

$c = \dfrac{73080 - 6(68)(178)}{28282 - 6(68)^2} = \dfrac{456}{538} = 0{,}85$

Met behulp van formule (10.8) vinden we dan: $d = 178 - 0{,}85(68) = 120{,}20$

zodat de vergelijking van de tweede regressielijn luidt: $\hat{x} = 0{,}85y + 120{,}20$.

Opdracht

Bereken voor de groep waaruit de 6 personen van voorbeeld 2 afkomstig zijn een schatting van de gemiddelde lichaamslengte van alle personen die een lichaamsgewicht hebben van 67 kg.

De figuren 10.4a en 10.4b laten duidelijk het onderscheid zien tussen de beide soorten regressielijnen.

Uit deze figuren blijkt dat de lijn $\hat{y} = ax + b$ in het algemeen een andere is dan de lijn $\hat{x} = cy + d$. Alleen wanneer alle r_i $(i = 1, 2, 3, ..., n)$ gelijk aan 0 zijn vallen de beide lijnen samen. Alle meetpunten liggen dan op de eerste én de tweede regressielijn. De vergelijking van de eerste regressielijn is dan om te schrijven naar die van de tweede regressielijn en omgekeerd.

Opmerking

a. Het punt met x-coördinaat $\bar{x}$ en y-coördinaat $\bar{y}$ ligt op beide regressielijnen (ga dit na door invulling). De lijnen snijden elkaar dus in dit punt.

b. Wanneer Y als de ingestelde (onafhankelijke) variabele wordt beschouwd en X als de afhankelijke, gemeten variabele, spreken we van regressie van X op Y.

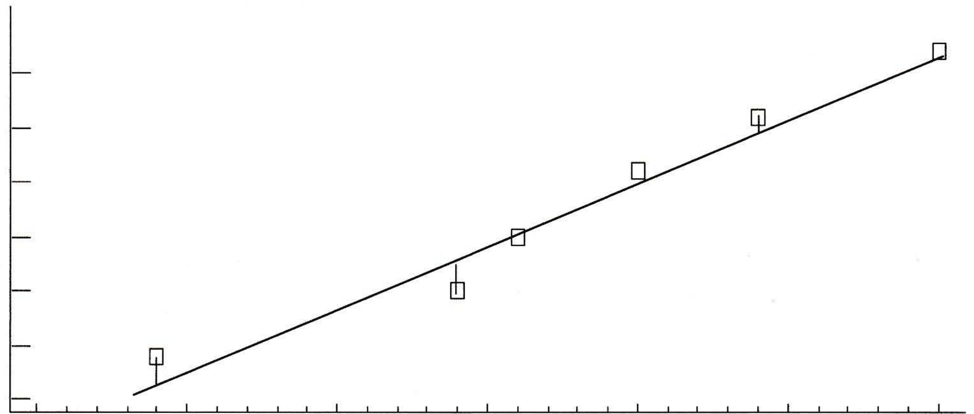


Fig. 10.4a De regressielijn $\hat{y} = 1{,}14x - 134{,}92$ van voorbeeld 1

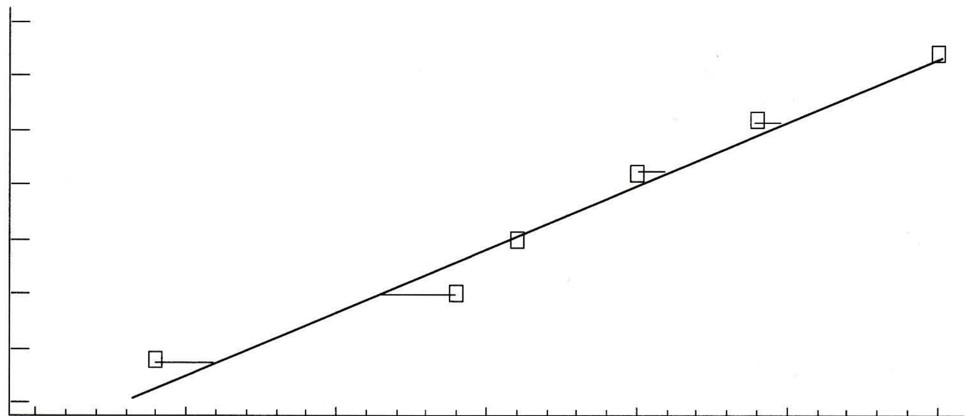


Fig. 10.4b De regressielijn $\hat{x} = 0{,}85y + 120{,}20$ van voorbeeld 2

10.4 Standaardfout

In de voorgaande paragrafen hebben we gezien hoe de eerste regressielijn $\hat{y} = ax + b$ en de tweede regressielijn $\hat{x} = cy + d$ bepaald konden worden op basis van n meetpunten (x_i, y_i). Bij de regressie van Y op X (Y gemeten, X onafhankelijk) is de som van de kwadraten van de residuen $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ een maat voor de schatting van de meetpunten door de eerste

regressielijn. Bij regressie van X op Y (X gemeten, Y onafhankelijk) is een maat voor de schatting van de meetpunten door de tweede regressielijn te geven door $\sum_{i=1}^{n}(x_i - \hat{x}_i)^2$.
We definiëren nu de *standaardfout* in de schatting van de meetpunten door de regressielijn als volgt.
Bij regressie van Y op X is de standaardfout in de schatting

$$s_{Y,X} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - ax_i - b)^2}{n}} \qquad (10.10)$$

terwijl deze bij regressie van X op Y gedefinieerd wordt als

$$s_{X,Y} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \hat{x}_i)^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - cy_i - d)^2}{n}} \qquad (10.11)$$

In het algemeen is $s_{Y,X}$ niet gelijk aan $s_{X,Y}$ (dit is verklaarbaar als we nog eens kijken naar figuur 10.4a en 10.4b).
De standaardfout heeft eigenschappen vergelijkbaar met die van de standaardafwijking. We herinneren eraan dat de afwijkingen van de meetwaarden bij een bepaalde ingestelde waarde ten opzichte van de regressielijn als normaal verdeel verondersteld worden met een gemiddelde 0 en een standaardafwijking s_r. Wanneer we evenwijdig aan de eerste regressielijn zowel erboven als eronder op een verticale afstand $s_{Y,X}$ lijnen zouden trekken, zal daarom blijken dat, zeker voor grote waarden van n, ongeveer 68% van de meetpunten tussen deze twee lijnen ligt. Immers: de kans dat een normaal verdeelde variabele een waarde bezit tussen $\mu - \sigma$ en $\mu + \sigma$ is $1 - 2P(U > 1) = 1 - (2)(0{,}1587) = 0{,}6826$.
Tussen twee evenwijdige lijnen op een afstand $2 \cdot s_{Y,X}$ boven en onder de eerste regressielijn ligt ongeveer 95% van alle meetpunten en tussen twee evenwijdige lijnen op een afstand $3 \cdot s_{Y,X}$ boven en onder de eerste regressielijn ligt 99,7% van alle meetpunten. Praktisch gezien liggen (vrijwel) alle meetwaarden y_i dus tussen $\hat{y}_i + 3 \cdot s_{Y,X}$ en $\hat{y}_i - 3 \cdot s_{Y,X}$.
Wanneer we slechts over weinig meetpunten beschikken, wordt een correctie op de formules (10.10) en (10.11) toegepast. We herinneren eraan dat bij het schatten van de standaardafwijking van een populatie in de formule voor de standaardafwijking van een steekproef door $n - 1$ gedeeld wordt in plaats van door n. Er gaat één vrijheidsgraad verloren door het schatten van het gemiddelde. Om die zelfde reden wordt bij de schatting van de standaardafwijking s_r in de noemer door $n - 2$ gedeeld. Er gaan door het schatten van a en b twee vrijheidsgraden verloren.

Conclusie:

$$s_{r,X} = s_{Y,X} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} \tag{10.12}$$

$$s_{r,Y} = s_{X,Y} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \hat{x}_i)^2}{n-2}} \tag{10.13}$$

Voorbeeld 3

Beschouw de zes meetpunten van voorbeeld 1.

nr	x_i	y_i	$\hat{y}_i = p(x_i) = 1{,}14x - 134{,}92$	$y_i - \hat{y}_i$
1	164	54	52,04	1,96
2	174	60	63,44	−3,44
3	176	65	65,72	−0,72
4	180	71	70,28	0,72
5	184	76	74,84	1,16
6	190	82	81,68	0,32

Met $n = 6$ en

$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = (1{,}96)^2 + (-3{,}44)^2 + (-0{,}72)^2 + (0{,}72)^2 + (1{,}16)^2 + (0{,}32)^2 = 18{,}16$

is $s_{Y,X} = \sqrt{\frac{18{,}16}{6-2}} = 2{,}1307$

Voor bijvoorbeeld een ingestelde waarde $x_i = 175$ cm vinden we voor de daarbij behorende meetwaarde y_i een spreidingsgebied

$$ax_i + b - 3s_{Y,X} < y_i < ax_i + b + 3s_{Y,X}$$

oftewel:

$$1{,}14 \cdot (175) - 134{,}92 - 3 \cdot (2{,}13) < y_i < 1{,}14 \cdot (175) - 134{,}92 + 3 \cdot (2{,}13)$$

dus:

$$58{,}19 < y_i < 70{,}97$$

Opdracht

Is het voor het in voorbeeld 1 beschreven geval aannemelijk dat iemand met een lengte van 180 cm 62 kg weegt? En dat hij 78 kg weegt? Wat is het laagste en wat is het hoogste aannemelijke lichaamsgewicht voor iemand die 180 cm lang is?

Bereken $s_{X,Y}$ en geef aan tussen welke waarden de lengte ligt van iemand met een gewicht van 60 kg

10.5 Niet-lineaire regressie

Soms zijn er duidelijk aanwijzingen dat het beter is door de puntenwolk van een aantal meetpunten (x_i, y_i) niet een rechte lijn te trekken, maar een kromme. Het verband tussen X en Y is dan niet lineair, maar in een aantal gevallen is het toch mogelijk het lineaire regressiemodel te gebruiken om de vergelijking van de niet-rechte regressiekromme te vinden. Stel bijvoorbeeld dat we een kromme van het type $y = p \cdot q^x$ willen trekken door een puntenwolk met n meetpunten (x_i, y_i) $(i = 1, 2, , n)$. Een dergelijke kromme zal men in de praktijk vaak tegenkomen bij groeiprocessen: Y is dan een maatstaf voor de groei (bijvoorbeeld van het aantal bacteriën in een kolonie) en X is een tijdvariabele.
Het (kromlijnige) model $y = p \cdot q^x$ kan getransformeerd worden naar een lineair model door aan beide kanten van de vergelijking $y = p \cdot q^x$ de logaritme te nemen (met welk grondtal dan ook).
We krijgen dan: $\log(y) = \log(p) + x\log(q)$, hetgeen met de substitutie $y' = \log(y)$ overgaat in $y' = \log(p) + x\log(q)$ en dus met $\log(p) = b$ en $\log(q) = a$ in $y' = ax + b$.
Met behulp van de ons bekende formules uit paragraaf 10.2 kunnen dan – door daarin y_i te vervangen door $y_i' = \log(y_i)$ – de coëfficiënten a en b berekend worden. Door middel van de terugtransformatie $p = g^b$ en $q = g^a$ (waarin g het grondtal van de gekozen logaritme voorstelt) vinden we dan de coëfficiënten p en q.
Het exponentiële model $y = p \cdot q^x$ is niet het enige niet-lineaire model dat in een lineair model getransformeerd kan worden. Andere voorbeelden zijn de modellen $y = p \cdot x^q$ en $y = p + q\cos x$.

Opdracht

Ga na tot welk lineair model het model $y = p + q\cos x$ getransformeerd kan worden en hoe de coëfficiënten p en q dan berekend kunnen worden.

Voorbeeld 4

Als toepassing van het niet-lineaire model $y = p \cdot x^q$ beschouwen we een zekere massa gas waarvoor bij 6 instelwaarden van het volume V en de druk P gemeten is. De resultaten waren als volgt:

nummer i	1	2	3	4	5	6
volume V	54,1	62,2	70,4	88,0	118,5	194,1
druk P	61,4	48,9	38,2	28,1	19,2	10,1

Volgens de thermodynamica bestaat de relatie $P \cdot V^k = C$ (ofwel $P = C \cdot V^{-k}$, een relatie van het type $y = p \cdot x^q$), waarin k en C constanten zijn. Bereken de meest waarschijnlijke waarden van k en C.

Oplossing

Omdat $P \cdot V^k = C$ is $\log P + k \log V = \log C$ (grondtal 10) hetgeen met $\log P = y$ en met $\log V = x$ overgaat in $y + k \cdot x = \log C$ en dus met $-k = a$ en met $\log C = b$ in $y = ax + b$.

Met $x_i = \log V_i$ en $y_i = \log P_i$ ($i = 1, 2, 3, ..., 6$) berekenen we nu analoog aan voorbeeld 1:

i	x_i	y_i	x_i^2	$x_i \cdot y_i$
1	1,7332	1,7882	3,0040	3,0993
2	1,7938	1,6893	3,2177	3,0303
3	1,8476	1,5821	3,4136	2,9231
4	1,9445	1,4487	3,7811	2,8170
5	2,0737	1,2833	4,3002	2,6612
6	2,2880	1,0043	5,2349	2,2978
totaal	11,6808	8,7959	22,9515	2,2978

Met $n = 6$, $\bar{x} = \dfrac{11{,}6808}{6} = 1{,}9468$ en $\bar{y} = \dfrac{8{,}7959}{6} = 1{,}4660$ vinden we:

$$a = \frac{16{,}8287 - 6(1{,}9468)(1{,}4660)}{22{,}9515 - 6(1{,}9468)^2} = \frac{0{,}2954}{0{,}2113} = 1{,}40 \text{ dus } k = -a = 1{,}40.$$

En verder: $b = 1{,}4660 - (-1{,}40)(1{,}9468) \approx 4{,}19$ dus $C = 10^b \approx 10^{4{,}19} = 15488$.
Het gezochte regressiemodel heeft dus de vorm $P \cdot V^{1{,}40} = 15488$.

10.6 Correlatierekening

10.6.1 De lineaire correlatiecoëfficiënt

Door de vergelijking van de lineaire regressielijn te vinden, kunnen we vaststellen van welke aard het stochastische ('toevallige') verband tussen twee variabelen is. Er zal altijd een lineair verband worden gevonden. Daarmee weten we echter nog niet direct hoe sterk dit verband is, met andere woorden hoe goed de gevonden regressielijn een weergave is van het eventuele verband tussen de beide variabelen. Een goede maatstaf voor het vastleggen van de mate van lineaire afhankelijkheid tussen twee kansvariabelen vinden we in de lineaire *correlatiecoëfficiënt*.

Voor een steekproef van n waarnemingsparen (x_i, y_i) ($i = 1, 2, ..., n$) wordt de lineaire correlatiecoëfficiënt aangeduid met het symbool $r(X, Y)$ en gedefinieerd als:

$$r(X,Y) = \frac{\sum_{i=1}^{n} x_i y_i - n \cdot \bar{x} \cdot \bar{y}}{\sqrt{\left(\sum_{i=1}^{n} (x_i)^2 - n \cdot (\bar{x})^2\right) \cdot \left(\sum_{i=1}^{n} (y_i)^2 - n \cdot (\bar{y})^2\right)}} \tag{10.14}$$

Een andere schrijfwijze voor deze formule is de volgende:

$$r(X,Y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\sum_{i=1}^{n} (x_i - \bar{x})^2\right) \cdot \left(\sum_{i=1}^{n} (y_i - \bar{y})^2\right)}} \tag{10.15}$$

Voorbeeld 5

Bereken volgens een van de formules (10.14) of (10.15) de lineaire correlatiecoëfficiënt $r(X, Y)$ voor de 6 meetpunten van voorbeeld 1.

Oplossing

Volgens formule (10.14) vinden we met $\bar{x} = 178$, $\bar{y} = 68$, $\sum_{i=1}^{6} x_i y_i = 73080$, $\sum_{i=1}^{6} (x_i)^2 = 190504$ en $\sum_{i=1}^{6} (y_i)^2 = 28282$:

$$r(X,Y) = \frac{73080 - 6(178)(68)}{\sqrt{(190504 - 6(178)^2)(28282 - 6(68)^2)}} = 0{,}983$$

Uit de formules (10.6) en (10.9) blijkt dat het kwadraat van de door formule (10.14) gedefinieerde lineaire correlatiecoëfficiënt $r(X, Y)$ gelijk is aan het product van de beide regressiecoëfficiënten a en c. Voor $r(X, Y)$ geldt dan:

$$r(X,Y) = \pm\sqrt{a \cdot c} \tag{10.16}$$

waarbij het plusteken geldt als a en c beide positief zijn en het minteken als a en c beide negatief zijn (het is niet mogelijk dat a en c een verschillend teken hebben).

Voorbeeld 6

Bereken volgens formule (10.16) de lineaire correlatiecoëfficiënt $r(X, Y)$ voor de 6 meetpunten van voorbeeld 1.

Oplossing

Met $a = 1{,}14$ en $c = 0{,}85$ vinden we volgens formule (10.16):
$r(X,Y) = \sqrt{1{,}14 \times 0{,}85} = \sqrt{0{,}969} = 0{,}984$.

Wanneer de beide regressielijnen $\hat{y} = ax+b$ en $\hat{x} = cy+d$ samenvallen, bestaat er tussen X en Y een lineair functioneel verband. Alle meetpunten liggen dan exact op de beide regressielijnen. Maar er geldt in dat geval ook dat de beide regressiecoëfficiënten a en c elkaars omgekeerde zijn, waardoor volgens formule (10.16) $r(X, Y) = -1$ (wanneer $a < 0$ en $c < 0$, dus wanneer de beide samenvallende regressielijnen dalend zijn) of $r(X, Y) = +1$ (wanneer $a > 0$ en $c > 0$, dus wanneer de beide samenvallende regressielijnen stijgend zijn). Met andere woorden: wanneer tussen twee variabelen X en Y een dalend respectievelijk stijgend lineair functioneel verband bestaat, dan is $r(X, Y) = -1$ respectievelijk $r(X, Y) = +1$.

Wanneer tussen twee kansvariabelen X en Y geen enkel lineair verband bestaat, is de lineaire correlatiecoëfficiënt $r(X, Y) = 0$. Dit betekent *niet* dat dan de variabelen X en Y per definitie onafhankelijk zijn: elk ander verband dan een lineair verband is in dat geval nog mogelijk.

Wanneer tussen twee kansvariabelen X en Y een lineair verband bestaat met een negatieve regressiecoëfficiënt, is $-1 < r(X, Y) < 0$. Hoe kleiner de spreiding van de punten rondom de regressielijn, hoe dichter $r(X, Y)$ bij -1 zal liggen; hoe groter die spreiding, hoe dichter $r(X, Y)$ bij 0 zal liggen. Is het verband tussen de beide variabelen echter lineair met een positieve regressiecoëfficiënt, dan is $0 < r(X, Y) < +1$ en zal $r(X, Y)$ bij een grotere spreiding van de punten rondom de regressielijn dichter bij 0 liggen en bij een kleinere spreiding dichter bij $+1$.

In figuur 10.5 zijn de eerder geschetste situaties nog eens aanschouwelijk voorgesteld.

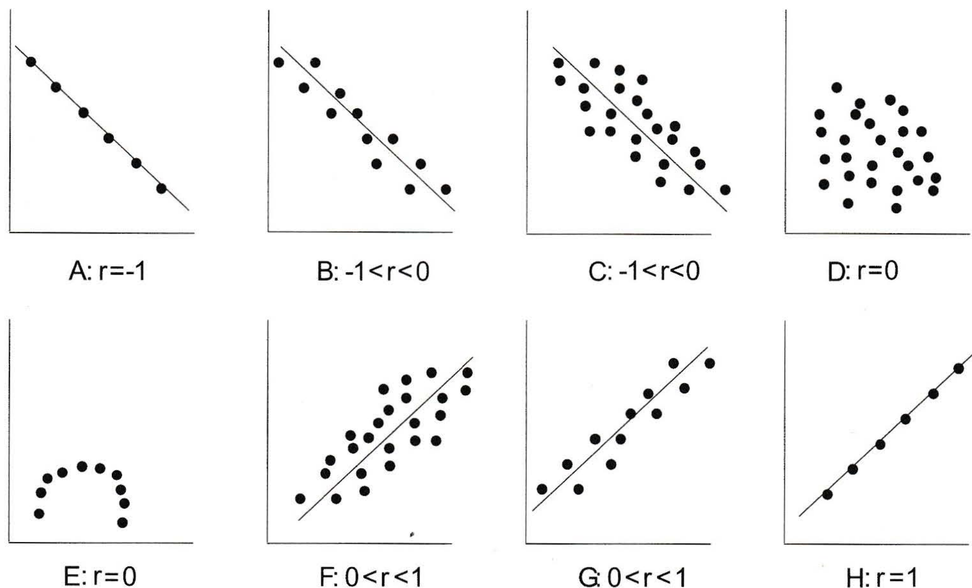


Fig. 10.5 Verschillende waarden van r

In geval B ligt $r(X, Y)$ dichter bij -1 dan in geval C. In geval G ligt r dichter bij $+1$ dan in geval F. In geval E is er wel een duidelijk verband maar dat is niet lineair. In dat geval geldt dus $r = 0$.

10.6.2 Het begrip covariantie

Voor een steekproef van n waarnemingsparen (x_i, y_i) $(i = 1, 2, 3, ..., n)$ wordt het begrip *covariantie* als volgt gedefinieerd:

$$cov(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n - 1} \tag{10.17}$$

of - wederom, omdat $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ te schrijven is als $\sum_{i=1}^{n}(x_i y_i) - n \cdot \bar{x} \cdot \bar{y}$:

$$cov(X, Y) = \frac{\sum_{i=1}^{n}(x_i y_i) - n \cdot \bar{x} \cdot \bar{y}}{n - 1} \tag{10.18}$$

Het directe gevolg van deze definitie is dat de zojuist gedefinieerde correlatiecoëfficiënt $r(X, Y)$ ook nog als volgt te schrijven is:

$$r(X, Y) = \frac{cov(X, Y)}{s_X \cdot s_Y} \tag{10.19}$$

waarbij s_X de standaardafwijking is van de x-coördinaten van de n meetpunten en s_Y de standaardafwijking is van de y-coördinaten van de n meetpunten. Merk op dat de benaming covariantie logisch is. Blijkbaar is deze grootheid te vergelijken met de variantie. Formule (10.19) gaat immers over in de formule voor de variantie van X wanneer tegelijk y_i vervangen wordt door x_i en $\bar{y}$ door $\bar{x}$.
Het ligt voor de hand om net als bij de *variantie* voor de covariantie van een *populatie* van coördinatenparen in de formules (10.17) en (10.18) niet door $n - 1$ maar door n te delen. Formule (10.19) blijft voor een populatie dezelfde, zij het dat de correlatiecoëfficiënt als parameter van de populatie (net als de standaardafwijking) met een Griekse letter geschreven wordt. Samengevat:

$$\rho(X, Y) = \frac{cov(X, Y)}{\sigma_X \cdot \sigma_Y} \tag{10.20}$$

met:

$$cov(X,Y) = \frac{\sum_{i=1}^{N}(x_i-\mu_X)(y_i-\mu_Y)}{N} \tag{10.21}$$

$$= \frac{\sum_{i=1}^{N}(x_i y_i) - N\cdot\mu_X\cdot\mu_Y}{N} \tag{10.22}$$

$$= \frac{\sum_{i=1}^{N}(x_i y_i)}{N} - \mu_X\cdot\mu_Y \tag{10.23}$$

Voorbeeld 7

In onderstaande tweedimensionale tabel is voor een populatie van 50 studenten vermeld welke examencijfers zij hadden voor de vakken wiskunde (X) en statistiek (Y).
Zo kan uit de tabel bijvoorbeeld worden afgelezen dat 7 studenten voor wiskunde een 6 hadden en voor statistiek een 7. Bereken de covariantie $cov(X,Y)$ en met behulp daarvan de lineaire correlatiecoëfficiënt.

y_i x_i	1	2	3	4	5	6	7	8	9	10	totaal
1	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	2	1	0	0	0	0	3
5	0	0	0	1	3	4	3	0	0	0	11
6	0	0	0	2	3	6	4	0	0	0	15
7	0	0	0	0	0	7	6	1	0	0	14
8	0	0	0	0	0	2	1	1	1	0	5
9	0	0	0	0	0	0	1	1	0	0	2
10	0	0	0	0	0	0	0	0	0	0	0
totaal	0	0	0	3	8	20	15	3	1	0	50

Oplossing

$$\mu_X = \frac{3(4)+8(5)+20(6)+15(7)+3(8)+1(9)}{50} = 6{,}2$$

$$\mu_Y = \frac{3(4)+11(5)+15(6)+14(7)+5(8)+2(9)}{50} = 6{,}26$$

$$\sum_{i=1}^{50} x_i y_i = 2(5\cdot 4)+1(6\cdot 4)+1(4\cdot 5)+3(5\cdot 5)+4(6\cdot 5)+3(7\cdot 5)+$$

$$= 2(4 \cdot 6) + 3(5 \cdot 6) + 6(6 \cdot 6) + 4(7 \cdot 6) + 7(6 \cdot 7) + 6(7 \cdot 7) + 1(8 \cdot 7) + 2(6 \cdot 8) + 1(7 \cdot 8) + 1(8 \cdot 8) + 1(9 \cdot 8) + 1(7 \cdot 9) + 1(8 \cdot 9)$$
$$= 1973$$

dus

$$cov(X, Y) = \frac{1973}{50} - (6{,}2)(6{,}26) = 0{,}648$$

$$\sum_{i=1}^{50} (x_i)^2 = 3(4^2) + 8(5^2) + 20(6^2) + 15(7^2) + 3(8^2) + 1(9^2) = 1976$$

Dus $\sigma_X^2 = \frac{1976}{50} - (6{,}2)^2 = 1{,}08$ en $\sigma_X = \sqrt{1{,}08} = 1{,}0392$

$$\sum_{i=1}^{50} (y_i)^2 = 3(4^2) + 11(5^2) + 15(6^2) + 14(7^2) + 5(8^2) + 2(9^2) = 2031$$

Dus $\sigma_Y^2 = \frac{2031}{50} - (6{,}26)^2 = 1{,}4324$ en $\sigma_Y = \sqrt{1{,}4324} = 1{,}1968$

Conclusie: $\rho(X, Y) = \frac{0{,}648}{(1{,}0392)(1{,}1968)} = 0{,}521$

In overeenstemming met formule (10.23) geldt voor de covariantie van twee kansvariabelen X en Y met verwachtingswaarden (gemiddelden) $E(X)$ en $E(Y)$:

$$cov(X, Y) = \frac{\sum_{i=1}^{N} (x_i y_i)}{N} - \mu_X \cdot \mu_Y = E(X \cdot Y) - E(X) \cdot E(Y) \qquad (10.24)$$

Wanneer twee kansvariabelen onafhankelijk zijn, zijn ze ook lineair onafhankelijk, dus is de lineaire correlatiecoëfficiënt gelijk aan 0. Volgens formule (10.20) is dan ook hun covariantie gelijk aan 0 en geldt er – zie formule (10.24) – dat $E(X \cdot Y) = E(X) \cdot E(Y)$.

10.7 Meervoudige regressie

Het kan voorkomen dat een variabele van twee of zelfs meer variabelen afhankelijk is. Wanneer deze afhankelijkheid lineair is kan als wiskundig model voor het verband tussen de variabelen Z, X en Y gebruikt worden de vergelijking:

$$z = a_0 + a_1 x + a_2 y \qquad (10.25)$$

Uit de wiskunde weten we dat formule (10.25) de vergelijking is van een plat vlak in de ruimte, die wordt opgespannen door een x-as, y-as en z-as. Wanneer we over een aantal (n) meetpunten met de coördinaten (x_i, y_i, z_i) beschikken, kunnen we zoeken naar het vlak dat

deze meetpunten het beste benadert. Door opnieuw het kleinste kwadratenkriterium toe te passen, komen we aan het *regressievlak*.
De coëfficiënten a_0, a_1 en a_2 die het regressievlak bepalen, vinden we door het volgende stelsel vergelijkingen op te lossen:

$$\begin{aligned}\sum_{i=1}^{n} z_i &= a_0 \cdot n + a_1 \cdot \sum_{i=1}^{n} x_i + a_2 \cdot \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} (x_i z_i) &= a_0 \cdot \sum_{i=1}^{n} x_i + a_1 \cdot \sum_{i=1}^{n} (x_i)^2 + a_2 \cdot \sum_{i=1}^{n} (x_i y_i) \\ \sum_{i=1}^{n} (y_i z_i) &= a_0 \cdot \sum_{i=1}^{n} y_i + a_1 \cdot \sum_{i=1}^{n} (x_i y_i) + a_2 \cdot \sum_{i=1}^{n} (y_i)^2\end{aligned} \qquad (10.26)$$

Om de mate van (lineaire) samenhang tussen de variabelen Z en X, Z en Y en ook X en Y te kunnen bepalen, kunnen we de formules voor de correlatiecoëfficiënt toepassen:

$$\rho_{Z,X} = \frac{\sum_{i=1}^{n} (z_i - \bar{z})(x_i - \bar{x})}{n \cdot \sigma_Z \cdot \sigma_X} \qquad (10.27)$$

$$\rho_{Z,Y} = \frac{\sum_{i=1}^{n} (z_i - \bar{z})(y_i - \bar{y})}{n \cdot \sigma_Z \cdot \sigma_Y} \qquad (10.28)$$

$$\rho_{Y,X} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{n \cdot \sigma_Y \cdot \sigma_X} \qquad (10.29)$$

waarbij:

$$\bar{z} = \frac{\sum_{i=1}^{n} z_i}{n} \quad \text{en} \quad \sigma_Z = \sqrt{\frac{\sum_{i=1}^{n} (z_i - \bar{z})^2}{n}}$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \quad \text{en} \quad \sigma_Z = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n}}$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{en} \quad \sigma_Z = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}$$

Bewezen kan worden dat de vergelijking van het regressievlak als volgt afhangt van de onderlinge correlatiecoëfficiënten:

$$\frac{z-\bar{z}}{\sigma_Z} = \left(\frac{\rho_{ZY} - \rho_{ZX} \cdot \rho_{YX}}{1-(\rho_{YX})^2}\right)\frac{y-\bar{y}}{\sigma_Y} + \left(\frac{\rho_{ZX} - \rho_{ZY} \cdot \rho_{YX}}{1-(\rho_{YX})^2}\right)\frac{x-\bar{x}}{\sigma_X} \qquad (10.30)$$

Voorbeeld 8

Een docent wil onderzoeken wat de invloed is op het tentamencijfer Z van een groep van 120 studenten van de resultaten van twee daaraan voorafgaande tussentoetsen (Y en X). De docent berekende daartoe de gemiddelde scores, de standaardafwijking en de onderlinge correlatiecoëfficiënten en kwam tot de volgende resultaten:

	Z	Y	X
gemiddelde	7,1	7,8	7,0
standaardafwijking	1,1	0,7	0,8

en correlatiecoëfficienten $\rho_{XY} = 0{,}6$, $\rho_{ZX} = 0{,}65$ en $\rho_{ZY} = 0{,}7$.

Voorbeeld 8 (vervolg)

De vergelijking van het regressievlak wordt

$$\frac{z-7{,}1}{1{,}1} = \left(\frac{0{,}7-0{,}65\times 0{,}6}{1-(0{,}6)^2}\right)\frac{y-7{,}8}{0{,}7} + \left(\frac{0{,}65-0{,}7\times 0{,}6}{1-(0{,}6)^2}\right)\frac{x-7{,}0}{0{,}8}$$

dus:

$$0{,}90909z - 6{,}4545 = 0{,}69196y + 0{,}44922x - 8{,}5419$$

oftewel:

$$z = 0{,}76116y + 0{,}49414x - 2{,}2961$$

Op basis van dit resultaat kan bijvoorbeeld geschat worden wat het tentamenresultaat is voor een student met een 9 voor de eerste tussentoets en een 7 voor de tweede tussentoets. $\hat{z} = 0{,}76116 \cdot 9 + 0{,}49414 \cdot 7 - 2{,}2961 = 8{,}0133$

Voor de standaardfout in het regressiemodel $z = a_0 + a_1x + a_2y$ (met Z als gemeten variabele) op basis van n meetpunten kan geschreven worden:

$$s_{Z,XY} = \sqrt{\frac{\sum_{i=1}^{n}(z_i-\hat{z}_i)^2}{n-3}} = \sqrt{\frac{\sum_{i=1}^{n}(z_i-(a_0+a_1x_i+a_2y_i))^2}{n-3}}$$

Het aantal vrijheidsgraden is hier $n-3$, omdat er drie coëfficiënten (a_0, a_1 en a_2) geschat moeten worden.

Meervoudige regressie

Wanneer de variabele Z van meer dan twee variabelen (lineair) afhankelijk is, is het regressiemodel nog verder uit te breiden. Voor de dan te gebruiken formules verwijzen we naar statistische literatuur. In statistische programmatuur (zoals ook in EXCEL) zijn dit soort modellen meestal wel aanwezig.

10.8 Het optellen en aftrekken van afhankelijke kansvariabelen

Als toepassing van het begrip *covariantie* keren we even terug naar hoofdstuk 7, waar we steeds variabelen bij elkaar opgeteld of van elkaar afgetrokken hebben met de aanname dat deze variabelen onafhankelijk zijn van elkaar. Bij het optellen en aftrekken van 2 variabelen die tot op zekere hoogte lineair van elkaar afhangen, mogen de formules (7.1) t/m (7.4) in aangepaste vorm gebruikt worden. De aanpassing betreft overigens uitsluitend de formule voor de variantie. Hierin wordt de afhankelijkheid van de twee variabelen als volgt betrokken.

Stelling 1

De som $Z = X + Y$ van twee normaal verdeelde kansvariabelen X en Y is normaal verdeeld.

Wat betreft het gemiddelde en de variantie van de som $Z = X + Y$ geldt:

$$\mu_Z = \mu_X + \mu_Y \qquad (10.31)$$

$$\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2 + 2 \cdot cov(X, Y) \qquad (10.32)$$

Stelling 2

Het verschil $Z = X - Y$ van twee normaal verdeelde kansvariabelen X en Y is normaal verdeeld. Wat betreft het gemiddelde en de variantie van het verschil $Z = X - Y$ geldt:

$$\mu_Z = \mu_X - \mu_Y \qquad (10.33)$$

$$\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2 - 2 \cdot cov(X, Y) \qquad (10.34)$$

Wanneer X en Y (lineair) onafhankelijk zijn, dus als de keuze van een waarde voor X totaal onafhankelijk is van de keuze van de waarde van Y, is de covariantie 0 en gaan de stellingen over in de stellingen 1 en 2 van hoofdstuk 7.

In het volgende voorbeeld geven we antwoord op een in paragraaf 7.2 gestelde vraag.

Voorbeeld 9

Uit een bak met staafjes (met normaal verdeelde lengte X, $\mu_X = 5$ cm en $\sigma_X = 0{,}2$ cm) pakt een robotarm een staafje en meet dit nauwkeurig op. Vervolgens wordt door

de robot net zo lang in de bak gezocht totdat deze een staafje vindt dat precies even lang is. Deze staafjes worden vervolgens aan elkaar gelast. Deze procedure wordt enige tijd herhaald. Bepaal de kansverdeling van de lengten van de aan elkaar gelaste staafjes.

Oplossing

Merk op dat hier niet aselect (onafhankelijk, stochastisch) wordt opgeteld. De lengte S van een gelast staafje is exact 2 maal de lengte van het eerst gekozen staafje: $S = 2 \cdot X$. Anders gezegd: de keuze van het tweede staafje is *volledig* (lineair) afhankelijk van de keuze van het eerste staafje en de correlatiecoëfficiënt van de beide variabelen X is dus 1.

Volgens de zojuist geformuleerde stelling geldt dan dat S normaal verdeeld is met: $\mu_S = \mu_X + \mu_X = 2 \cdot \mu_X = 10$ cm en
$\sigma_S^2 = \sigma_X^2 + \sigma_X^2 + 2 \cdot cov(X, X) = \sigma_X^2 + \sigma_X^2 + 2 \cdot \rho_{X,X} \cdot \sigma_X \cdot \sigma_X = 4 \cdot \sigma_X^2$ zodat $\sigma_S = 2 \cdot \sigma_X = 0{,}4$ cm.

We merken op dat dit resultaat geheel in overeenstemming is met wat we in hoofdstuk 3 geleerd hebben: wanneer alle waarnemingsuitkomsten met 2 vermenigvuldigd worden, worden gemiddelde en standaardafwijking eveneens met 2 vermenigvuldigd.

Pas op: wanneer de keuze van het tweede staafje steeds willekeurig (aselect) was geweest, hadden we wel hetzelfde gemiddelde maar niet dezelfde standaardafwijking gekregen!! Immers, in dat geval ($S = X + X$), zou volgens stelling 1 uit hoofdstuk 7 gelden: $\sigma_S^2 = \sigma_X^2 + \sigma_X^2 = 2 \cdot \sigma_X^2$ zodat $\sigma_S = \sigma_X \cdot \sqrt{2}$.

Opdracht

Geef een verklaring (zonder formules) waarom in het laatste geval de standaardafwijking kleiner is dan in het voorbeeld zelf.

Voorbeeld 10

Een project bestaat uit de onderling onafhankelijke activiteiten A, B, C, D en E, die direct na elkaar worden uitgevoerd. De activiteiten hebben een tijdsduur die normaal verdeeld is. Gemiddelde en standaardafwijking staan in de tabel:

activiteit	μ (in dagen)	σ (in dagen)
A	2,1	0,2
B	3,4	0,4
C	4,1	0,5
D	5,2	0,3
E	3,2	0,4

Welke invloed heeft activiteit E op de totale projectduur?

Oplossing

Het is vanzelfsprekend dat activiteit E (maar ook de overige activiteiten) invloed heeft op de totale projectduur. Voor de totale projectduur T geldt: $T = A+B+C+D+E$, waarbij we de letters A, B, C, D en E gebruikt hebben voor de tijdsduur van de betreffende activiteiten. Omdat A t/m E onafhankelijk van elkaar zijn, kunnen we (volgens stelling 5 uit hoofdstuk 7) stellen dat T normaal verdeeld is met gemiddelde $\mu_T = \mu_A + \mu_B + \mu_C + \mu_D + \mu_E = 18$ dagen en variantie $\sigma_T^2 = \sigma_A^2 + \sigma_B^2 + \sigma_C^2 + \sigma_D^2 + \sigma_E^2 = 0{,}7$ dus standaardafwijking $\sigma_T = 0{,}837$ dagen.

Om de mate van afhankelijkheid te kunnen bepalen tussen T en E, berekenen we de correlatiecoëfficiënt met behulp van de covariantie $cov(T, E)$. Deze covariantie kunnen we berekenen door te bedenken dat $A + B + C + D = T - E$.

De som $A + B + C + D$ is normaal verdeeld met (volgens stelling 5 uit hoofdstuk 7) gemiddelde $\mu_{A+B+C+D} = \mu_A + \mu_B + \mu_C + \mu_D = 14{,}8$ dagen en variantie $\sigma^2_{A+B+C+D} = \sigma_A^2 + \sigma_B^2 + \sigma_C^2 + \sigma_D^2 = 0{,}54$ dus $\sigma_{A+B+C+D} = 0{,}735$ dagen.

Formule (10.34) leert nu dat $\sigma^2_{A+B+C+D} = \sigma_T^2 + \sigma_E^2 - 2 \cdot cov(T, E)$ dus $0{,}54 = 0{,}7 + 0{,}16 - 2 \cdot cov(T, E)$, zodat $cov(T, E) = \dfrac{0{,}54 - 0{,}86}{-2} = 0{,}16$.

De correlatiecoëfficiënt tussen de totale projectduur T en de duur van activiteit E is dan volgens formule (10.20):

$$\rho_{T,E} = \frac{cov(T, E)}{\sigma_T \cdot \sigma_B} = \frac{0{,}16}{(0{,}837)(0{,}4)} = 0{,}48.$$

Opgaven

1. Gegeven zijn de volgende waarnemingsparen:

i	1	2	3	4	5
x_i	3	-2	4	0	5
y_i	7	-6	16	0	13

Bepaal de vergelijking van de eerste regressielijn en bereken de som van de residuen ten opzichte van deze lijn. Bereken ook de som van de kwadraten van de residuen.

2. Gegeven zijn de volgende waarnemingsparen:

i	1	2	3	4	5	6	7	8
x_i	15	21	24	10	6	18	20	14
y_i	6,0	9,0	10,5	3,5	1,5	7,5	8,5	5,5

Bepaal de vergelijking van de tweede regressielijn en bereken de som van de residuen ten opzichte van deze lijn. Bereken ook de som van de kwadraten van de residuen.

3. Gegeven zijn de volgende waarnemingsparen:

i	1	2	3	4	5	6	7	8	9	10
x_i	0,9	0,2	0,6	0,8	0,4	0,7	1,0	0,1	0,3	0,5
y_i	17	4	10	14	18	7	22	21	19	23

a. Bepaal de regressielijn van Y op X.
b. Bepaal de regressielijn van X op Y.
c. Bepaal de correlatiecoëfficiënt tussen X en Y.

4. Op een machine worden ronde aluminium staven met een lengte van 100 cm in stukjes van 10 cm gezaagd. Hoe groter de diameter van de staaf, hoe groter de zaagtijd is. Om te onderzoeken wat het verband is tussen diameter en zaagtijd, werd van 9 staven de diameter (D) gemeten, waarna de 9 staven een voor een in stukjes werden gezaagd en voor elke staaf de zaagtijd (Z) werd gemeten. De resultaten waren als volgt:

i	1	2	3	4	5	6	7	8	9
d_i	12,1	12,7	11,4	13,6	12,0	15,4	15,9	14,7	15,8
z_i	20,8	21,2	19,6	23,6	21,0	26,8	28,0	25,6	27,8

a. Bepaal de vergelijkingen van de eerste en de tweede regressielijn.
b. Bereken de correlatiecoëfficiënt tussen de zaagtijd en de diameter.
c. Voor een tiende staaf met een diameter van 11,7 mm werd een zaagtijd van 21,4 seconden gemeten. Is er reden om deze tijdmeting te wantrouwen?

5. Om te onderzoeken of het aantal in rollen gordijnstof voorkomende weeffouten des te groter is naarmate de rollen langer zijn, werd van 14 rollen van verschillende lengte het in de rollen voorkomende aantal weeffouten geteld.
De resultaten waren als volgt:

nr.rol	lengte in m	aantal weeffouten	nr. rol	lengte in m	aantal weeffouten
1	160	2	8	480	3
2	240	2	9	580	5
3	80	3	10	580	7
4	240	2	11	580	7
5	260	6	12	200	3
6	460	4	13	580	5
7	480	3	14	720	8

a. Bepaal de vergelijkingen van de eerste en de tweede regressielijn.
b. Bepaal de correlatiecoëfficiënt tussen het aantal weeffouten per rol en de lengte van de rollen.

6. Op een consultatiebureau wordt voortdurend onderzocht wat de relatie is tussen leeftijd (variabele Z), lengte (variabele X) en gewicht (variabele Y) van jonge kinderen. Bij een kind werden de volgende metingen gedaan:

meting	1	2	3	4	5	6
leeftijd (in dagen)	466	564	863	915	1091	1460
lengte (in cm)	78	82	91	95	101	109
gewicht (in pond)	10,1	11,6	14	15,1	16,6	19,6

Bepaal een lineair regressiemodel op grond van de gegevens.

7. Het aantal bacteriën per volume-eenheid in een bacteriëncultuur groeit exponentieel in de tijd. Als Y het aantal bacteriën per volume-eenheid is en X de tijd in uren, geldt er $y = a \cdot b^x$. In een zeker onderzoek werden de volgende resultaten gevonden:

i	1	2	3	4	5	6	7
x_i	0	1	2	3	4	5	6
y_i	30	45	63	91	132	191	278

Bereken, gebruikmakend van de methode van de kleinste-kwadraten, de meest waarschijnlijke waarde van a en b.

8. Gegeven zijn de volgende waarnemingsparen:

i	1	2	3	4	5	6
x_i	1	5	6	8	9	10
y_i	1	5	11	13	20	26

a. Teken van deze 6 meetpunten een puntendiagram.
b. Teken in het puntendiagram een kromme van het type $y = a + b \cdot x^2$.
c. Dezelfde vraag voor een kromme van het type $y = a \cdot x^b$.

9. De scores van een grote groep studenten voor een tentamen wiskunde bleken normaal verdeeld te zijn met gemiddelde $\mu_W = 6{,}35$ en standaardafwijking $\sigma_W = 0{,}9$. De scores van dezelfde groep studenten voor een tentamen statistiek bleken eveneens normaal verdeeld te zijn, echter met gemiddelde $\mu_S = 5{,}90$ en standaardafwijking $\sigma_S = 1{,}20$. Voor 33% van de studenten lag de score voor het tentamen wiskunde meer dan 1 punt boven de score voor het tentamen statistiek. Bereken de correlatiecoëfficiënt tussen de scores voor de beide vakken.

10. In een audioversterker worden twee transistors van hetzelfde type in serie geschakeld, waarbij de totale versterking tussen 40 dB en 60 dB moet liggen. De transistors worden geleverd in partijen waarvan de versterking normaal verdeeld is met een gemiddelde van 25 dB en een standaardafwijking van $3\sqrt{2}$ dB.
De transistors worden bij ontvangst gesplitst in transistors met een versterking die lager is dan het gemiddelde (A-transistors) en transistors met een versterking die hoger is dan het gemiddelde (B-transistors).
Daardoor ontstaan partijen A-transistors met een gemiddelde versterking van 20 dB en een standaardafwijking van 3 dB (normaal verdeeld) en partijen B-transistors met een gemiddelde versterking van 30 dB en een standaardafwijking van 4 dB (eveneens normaal verdeeld).

a. Wanneer men in de versterkers twee transistors inbouwt die willekeurig uit een niet-gesplitste partij worden betrokken, hoeveel procent uitval zal dan ten aanzien van de versterking optreden?
b. Hoe hoog zal dit percentage zijn wanneer men een A-transistor en een B-transistor willekeurig combineert en inbouwt?
c. Wanneer men met behulp van bepaalde apparatuur bij iedere A-transistor een B-transistor voegt, zodanig dat slechts 1,24% van de ontstane combinaties niet aan de gestelde versterkingseis voldoet (0,62% te laag en 0,62% te hoog), hoe groot is dan de correlatiecoëfficiënt tussen de versterking van de beide transistors?
d. Wanneer de correlatiecoëfficiënt tussen de versterking van de beide transistors $\rho = -\frac{2}{3}$ bedraagt, hoeveel procent van de combinaties voldoet dan niet aan de gestelde versterkingseis?
e. Hoe groot is in het geval dat men een A-transistor en een B-transistor willekeurig combineert en inbouwt, de correlatiecoëfficiënt tussen de versterking van de A-transistors en die van de combinaties?
f. Hoe groot is de correlatiecoëfficiënt tussen de versterking van de B-transistors en die van de combinaties?

11 Statistische procesbeheersing

11.1 Inleiding

Het thema 'statistische procescontrole' SPC (Engels: *statistical process control*) staat de laatste tijd sterk in de belangstelling. We kunnen spreken van een deels hernieuwde belangstelling want de grondbeginselen van SPC werden al in de jaren 1920-1930 geïntroduceerd door pioniers als Shewhart (USA) en Tippett (GB).
Aan de doelstelling van SPC en gebruik van statistische technieken zijn in latere jaren, beginnend rond 1960, nieuwe elementen toegevoegd: pioniers zijn onder andere Deming en Juran.
Bij Shewhart ligt de directe doelstelling in het statistisch beheerst maken en houden van een proces, Juran en vooral Deming voegen daaraan toe: de continue verbetering van het proces en daarmee de continue verbetering van de productkwaliteit.
In verband met beide doelstellingen gebruikt men dezelfde statistische basisinformatie van het proces.
De procesbeheersing is vooral een opgave voor de mensen die in het productieproces zelf actief zijn.
Hetgeen zojuist is gezegd, wordt nog eens toegelicht aan de hand van figuur 11.1.

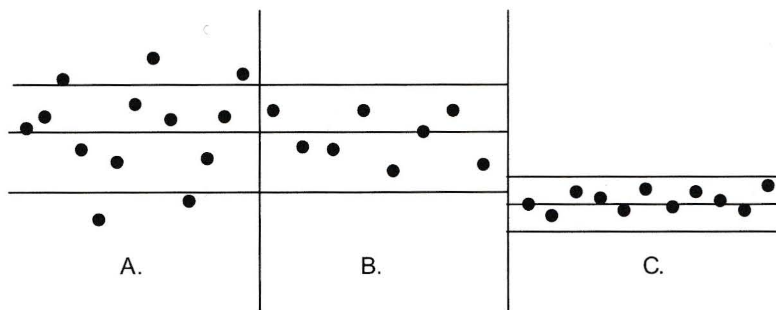


Fig. 11.1 Controlekaart van een proces

In figuur 11.1 is de controlekaart van een proces weergegeven, gezien in de ontwikkeling van de tijd.

Fase A: *Het proces is niet onder statistische controle, met als gevolgen:*
- de kwaliteit van het product is onzeker;
- er zijn relatief hoge kosten voor inspectie en correctie.

Fase B: *Het proces is onder statistische controle gebracht*
- de kwaliteit van het product spreidt alleen ten gevolge van *procesinherente* spreiding;
- de productkwaliteit is voorspelbaar.

Fase C: *In het proces zijn wijzigingen aangebracht*
- waardoor de productkwaliteit qua gemiddelde en spreiding op een gunstiger niveau is gebracht.

Definitie

SPC is de inzet van een groot aantal (statistische) methoden om variaties in de kwaliteit- en procesgegevens aan het licht te brengen met het doel maatregelen te kunnen treffen om een gelijkmatiger en zich continu verbeterende productkwaliteit te verkrijgen.

Naarmate de productkwaliteit zich verbetert, zal SPC ook tot een toename van productiviteit leiden, tot een daling van het energieverbruik en tot een verhoging van de concurrentiekracht.
De reeds hiervoor genoemde W. Edwards Deming, die SPC op een breed gebied in de Japanse en (later!) de Amerikaanse industrie invoerde, definieert SPC als volgt:

'SPC is the application of statistical principles and techniques in all stages of production directed towards the economic manufacture of a product that is useful and has a market.'

Doelstellingen van SPC zijn samengevat:
- het proces onder statistische controle brengen en houden;
- het proces en het product verbeteren;
- het zwaartepunt verleggen van productinspectie naar procesbeheersing.

11.2 Controlekaarten

In dit hoofdstuk gaan we verder in op de statistische procesbeheersing. Een belangrijk hulpmiddel hierbij is de *controlekaart*.
Het doel van het gebruik van controlekaarten is tweeledig.

a. Het proces wordt door middel van een controlekaart continu bewaakt en geëvalueerd, teneinde het proces statistisch beheerst te houden. Een belangrijke taak hierbij is vast te stellen of het proces veranderd is (*assignable cause*) en bijgesteld moet worden.

b. Indien het proces *niet* statistisch beheerst is, helpt een controlekaart je het onder 'statistical control' te brengen.

Een controlekaart wordt meestal aangelegd voor het gemiddelde ($\bar{x}$) en de standaardafwijking (s) van de steekproef. Aan de hand van ($\bar{x}$) kan men nagaan of het werkelijke proces verschoven is, dus of het (onbekende) populatiegemiddelde μ_i afwijkt van de 'target' μ_0. Daartoe worden regelmatig steekproeven uit het lopend proces genomen en in een grafiek ingetekend (zie fig. 11.2).

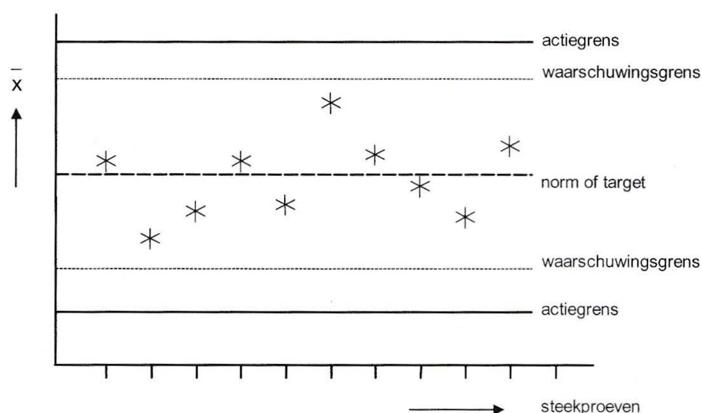


Fig. 11.2 Voorbeeld controlekaart voor $\bar{x}$

Bij een statistisch beheerst proces zal het steekproefgemiddelde ($\bar{x}$) zich bewegen rond een constant procesniveau μ_0. De afwijkingen van $\bar{x}$ ten opzichte van μ_0 zijn het resultaat van de som van op zichzelf kleine effecten van tal van factoren die op het proces inwerken. Voorbeelden zijn: 'normale' temperatuurfluctuaties, variaties in grondstof, enzovoorts.
In figuur 11.3 is een controlekaart gegeven van een proces dat statistisch beheerst verloopt tot steekproefnummer 12, waarna het proces systematisch gaat afwijken van μ_0.

11.3 Doel en opzetten van verschillende typen controlekaarten

Het type controlekaart dat in de inleiding is beschreven, is de Shewhart-controlekaart, genoemd naar de Amerikaanse statisticus Shewhart die deze controlekaart in 1924 introduceerde.
In het voorbeeld is uitgegaan van continue kwantitatieve kenmerken, zoals sterkte, viscositeit, enzovoorts. Deze controlekaarten zijn gebaseerd op de normale kansverdeling. Dezelfde statistische principes kunnen we toepassen bij discrete kenmerken, zoals het percentage defecten in een partij, het aantal breuken, het aantal storingen, enzovoorts. Voor dergelijke kenmerken zijn de binomiale en Poisson-kansverdelingen van toepassing.

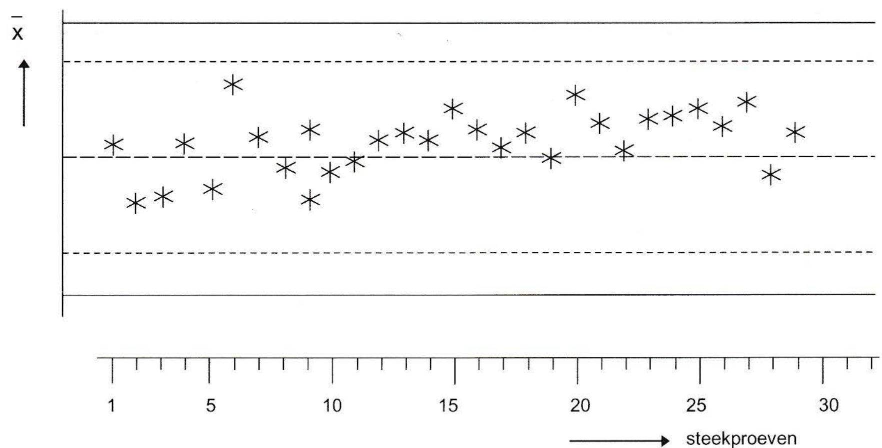


Fig. 11.3 Voorbeeld controlekaart van proces waarin het procesniveau is veranderd

Andere vormen van controlekaarten zijn:
- *de controlekaart voor individuen*, toegepast indien slechts één waarneming per steekproef wordt verricht;
- de *goed- of afkeurkaart*, waarbij bepaalde fluctuaties in niveau zijn toegestaan, echter gereageerd moet worden indien bepaalde toleranties dreigen te worden overschreden;

11.3.1 De Shewhart-controlekaart voor kwantitatief meetbare eigenschappen

Het doel van de *Shewhart-controlekaart* is voldoende besproken in de inleiding van dit hoofdstuk. In deze paragraaf zullen we de aandacht richten op het inrichten van een Shewhartkaart. Belangrijke uitgangspunten bij de Shewhart-controlekaart voor kwantitatief meetbare eigenschappen zijn:
- de toevallige afwijkingen kunnen praktisch voldoende nauwkeurig door de 'normale verdeling' worden beschreven;
- de meetresultaten van de steekproeven zijn onderling onafhankelijk (er is geen correlatie).

De berekening van de grenzen in de Shewhart-controlekaart is gebaseerd op informatie van de hiervoor genoemde normale verdeling van de steekproefuitkomsten als het proces statistisch beheerst verloopt. In het onderstaande zullen we ingaan op de berekening van deze grenzen.
Een beslissingsregel bij de Shewhart-controlekaart is onder andere:

Als een waarneming buiten de actiegrenzen valt, dient de oorzaak van de afwijking opgespoord te worden en gecorrigeerd.

11.3.2 Het berekenen van de grenzen in de Shewhart-controlekaart

Hierbij worden twee gevallen onderscheiden:

1. populatiegemiddelde en spreiding (μ en σ) zijn bekend;
2. populatiegemiddelde en spreiding zijn niet bekend.

***1.* Gemiddelde en spreiding zijn bekend (*μ en σ zijn bekend*)**

In het geval dat het populatiegemiddelde μ en de populatiespreiding σ bekend zijn, is het zo dat het gemeten kenmerk afkomstig is van productie-eenheden over lange tijd, uit een ongestoord verlopen proces. We kunnen nu de normwaarde en de grenzen voor een controlekaart berekenen voor steekproeven van n stuks.

a. $\bar{x}$-kaart

normlijn: populatiegemiddelde μ	
Bovenste actiegrens:	$\mu + 3\sigma_{\bar{X}} = \mu + 3\frac{\sigma}{\sqrt{n}}$
Onderste actiegrens:	$\mu - 3\sigma_{\bar{X}} = \mu - 3\frac{\sigma}{\sqrt{n}}$

b. R-kaart

Bij het berekenen van de bovengrens voor de R-kaart hebben we enkele nieuwe factoren nodig. Voor de gemiddelde range maken we gebruik van de factor d_2, voor de bovengrens van D_2 en voor de ondergrens van D_1. De benodigde factoren zijn vermeld in tabel B12. Deze factoren zijn afhankelijk van de steekproefgrootte en niet van het aantal steekproeven. Met behulp hiervan kunnen we de lijnen op de R-kaart als volgt berekenen:

Target: gemiddelde range $d_2 \cdot \sigma$	
Onderste actiegrens:	$D_1 \cdot \sigma$
Bovenste actiegrens:	$D_2 \cdot \sigma$

c. s- kaart

Indien een spreidingskaart voor de steekproefstandaardafwijking s moet worden opgesteld, gaat dit als volgt (ook de hiervoor benodigde constanten zijn in tabel B12 vermeld):

Target: $c_4 \cdot \sigma$	
Onderste actiegrens:	$B_1 \cdot \sigma$
Bovenste actiegrens:	$B_2 \cdot \sigma$

2. Gemiddelde en spreiding zijn onbekend

In het geval dat zowel populatiegemiddelde als spreiding niet bekend zijn, moeten we deze eerst schatten. Dit doen we aan de hand van de resultaten van een voldoend groot aantal steekproeven (dit aantal ligt in de praktijk meestal rond de 20) met dezelfde omvang als die waarvoor we de controlekaarten willen gaan gebruiken.
Uit deze gegevens berekenen we per steekproef het gemiddelde $\bar{x}$ en de range (R) en daarna het gemiddelde van alle steekproefgemiddelden ($\bar{x}_{gem}$) en het gemiddelde van de ranges ($\bar{R}$). Wel dient nog opgemerkt te worden dat de circa 20 steekproeven die nodig waren, niet direct na elkaar genomen mogen worden, maar verspreid in de tijd en dan nog uit een periode waarvan men overtuigd is, dat er geen storingen in het proces zijn opgetreden.
De benodigde gegevens voor de $\bar{x}$-kaart en de R-kaart berekenen we nu als volgt:

a. $\bar{x}$-kaart

Target: gemiddelde van alle steekproefgemiddelden $\bar{x}_{gem}$	
Onderste actiegrens:	$\bar{x}_{gem} - A_2 \cdot \bar{R}$
Bovenste actiegrens:	$\bar{x}_{gem} + A_2 \cdot \bar{R}$

De factor A_2 staat vermeld in tabel B12.

b. R-kaart

Target: gemiddelde range $\bar{R}$	
Onderste actiegrens:	$D_3 \cdot \bar{R}$
Bovenste actiegrens:	$D_4 \cdot \bar{R}$

De factoren D_3 en D_4 staan wederom vermeld in tabel B12.

c. Analoog volgt voor de $\bar{x}$- en s-kaart:

$\bar{x}$-kaart

Target: totaal gemiddelde $\bar{x}_{gem}$	
Onderste actiegrens:	$\bar{x}_{gem} - A_3 \cdot \bar{s}$
Bovenste actiegrens:	$\bar{x}_{gem} + A_3 \cdot \bar{s}$

s-kaart

Target: $\bar{s}$	
Onderste actiegrens:	$D_3 \cdot \bar{s}$
Bovenste actiegrens:	$D_4 \cdot \bar{s}$

Voorbeeld 1

Als voorbeeld voor het opstellen van een $\bar{x}$- en R- kaart nemen we de gegevens in onderstaande tabel. In deze tabel staan de resultaten vermeld van 20 steekproeven van $n = 5$ stuks, waarbij het meetkenmerk het gewicht van een bepaald product is. Het populatiegemiddelde (μ) en de populatiespreiding (σ) zijn in dit geval onbekend en zullen dus geschat moeten worden uit de genomen steekproeven.

steekpr. nr	meetuitkomsten					$\bar{x}$	R
1	467,8	465,5	468,1	468,9	468,4	467,7	3,4
2	467,5	465,2	468,4	469,3	468,8	467,8	4,1
3	467,3	468,6	467,6	471,2	468,1	468,6	3,9
4	470,5	467,9	470,4	477,0	468,7	470,9	9,1
5	465,6	469,0	467,7	468,8	470,8	468,4	5,2
6	467,0	469,2	469,9	468,4	469,0	468,7	2,9
7	467,8	470,2	464,8	464,6	469,0	467,3	5,6
8	467,4	471,2	467,2	467,0	466,2	467,8	5,0
9	466,2	471,8	467,2	468,5	466,2	468,0	5,6
10	468,9	469,3	466,6	467,5	467,2	467,9	2,7
11	464,6	469,9	468,7	468,2	468,0	467,9	5,3
12	472,6	468,8	467,6	469,5	469,7	469,6	5,0
13	471,1	468,8	467,4	465,1	467,1	467,9	6,0
14	467,1	467,0	470,8	466,1	470,5	468,3	4,7
15	469,4	469,3	467,4	469,0	471,7	469,4	4,3
16	467,4	468,3	466,9	469,7	468,2	468,1	2,8
17	469,8	471,7	468,7	471,0	472,5	470,7	3,7
18	467,4	463,2	468,8	470,6	466,8	467,4	7,4
19	465,9	461,6	468,6	465,0	469,2	466,1	7,6
20	463,6	465,8	469,3	467,3	470,3	467,3	6,7

Tabel 11.1 Meetwaarden van 20 steekproeven ($n = 5$)

1. Bepalen van controlegrenzen voor de $\bar{x}$-kaart:
Van elke steekproef wordt eerst het gemiddelde ($\bar{x}$) en de range (R) bepaald. Vervolgens worden de lijnen van de controlekaart berekend: Normlijn of target is het overall gemiddelde: $\bar{x}_{gem}$ (= 468,29).
Voor het bepalen van de boven- en ondergrens wordt eerst de gemiddelde range ($\bar{R}$) bepaald (= 5,06):
De bovengrens is: $\bar{x}_{gem} + A_2 \cdot \bar{R} = 468{,}29 + 0{,}577 \cdot 5{,}06 = 471{,}21$
De ondergrens is: $\bar{x}_{gem} - A_2 \cdot \bar{R} = 468{,}29 - 0{,}577 \cdot 5{,}06 = 465{,}36$
De factor A_2 is opgezocht in de tabel bij een steekproefgrootte $n = 5$: $A_2 = 0{,}577$.

2. Bepalen van de controlegrenzen voor de R-kaart
Gemiddelde range: $\bar{R} = 5{,}06$
Bovengrens: $D_4 \cdot \bar{R}$: $2{,}115 \cdot 5{,}06 = 10{,}70$
Ondergrens: $D_3 \cdot \bar{R}$: $0 \cdot 5{,}06 = 0$
De controlekaart uit dit voorbeeld is gegeven in figuur 11.4.

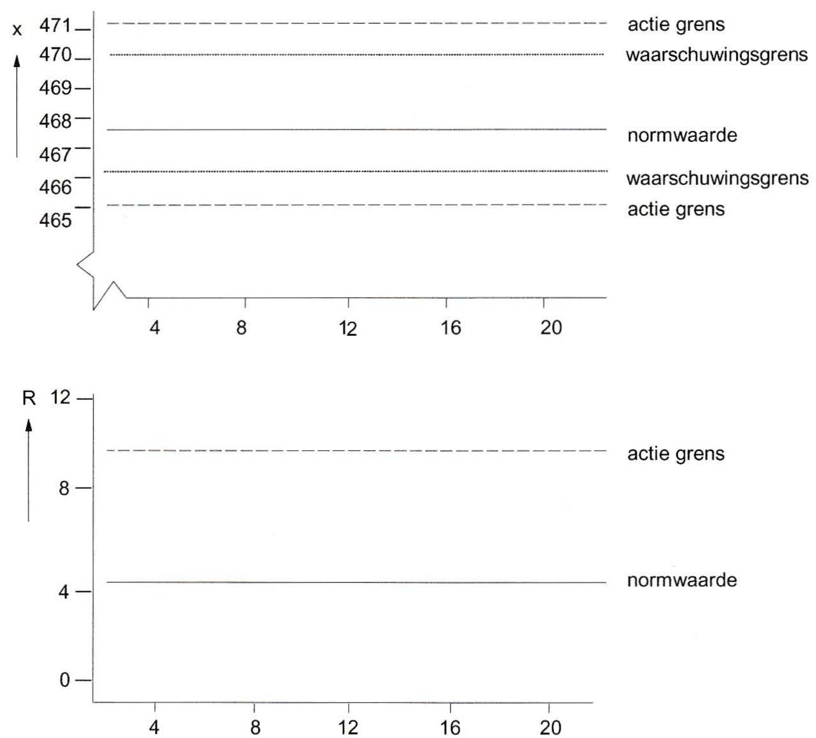


Fig. 11.4 $\bar{x}$ en R- kaart voor de gegevens van tabel 11.1

11.4 Controlekaart voor individuen

In veel situaties is het niet goed mogelijk om een proces op te splitsen in subgroepen van individuen, welke men vervolgens via een Shewhart-controlekaart kan bewaken. Dit is met name het geval in batchprocessen waarbij per batch (=serie) een homogene partij wordt geproduceerd. Neemt men van een dergelijke partij meerdere monsters, dan zullen doorgaans tussen deze monsters relatief kleine verschillen optreden. De voor controle relevante variaties zullen echter optreden tussen de batches onderling. Men kan in dit verband gebruik maken van 'moving ranges'. Hierbij wordt de range (R) van 2 of meer (algemeen n) batches bepaald, het gemiddelde daarvan is $\bar{R}$. Uit $\bar{R}$ kan nu een schatting van de spreiding (σ) van het proces worden bepaald via:

$$\hat{\sigma} = \frac{\bar{R}}{d_2} \qquad (11.1)$$

De waarde van d_2 is afhankelijk van het aantal opeenvolgende batches (n) waaruit $\bar{R}$ wordt geschat, en is in tabel B12 gegeven.
De controlegrenzen worden berekend, gebruikmakend van de hierboven besproken schatting van de standaardafwijking $\hat{\sigma}$ uit de moving ranges (R).

Voorbeeld 2

In Tabel 11.2 is een voorbeeld gegeven van de berekening van de moving range, gebaseerd op de verschillen van twee opeenvolgende batches. In het voorbeeld wordt de relatieve viscositeit bepaald van de productie van een product per batch.

batch	rel. viscositeit	moving range (R)
1	4,3	-
2	4,6	0,3
3	4,9	0,3
4	4,5	0,4
5	5,0	0,5
6	4,7	0,3
7	4,7	0
8	4,9	0,2
9	4,6	0,3
10	4,7	0,1
$\sum$	46,9	2,4

Tabel 11.2 Voorbeeld moving ranges

Oplossing

$d_2 = 1{,}128$ ($n = 2$), zie tabel B12.

$\hat{\sigma} = \dfrac{0{,}27}{1{,}128} = 0{,}24$

Deze waarde gebruiken we voor de berekening van de controlelijnen in de $\bar{x}$- R- kaart.

Voor meting per batch:

target: $\bar{x} = \dfrac{46{,}9}{10} = 4{,}69$

Bovenste actiegrens: $\bar{x} + 3\hat{\sigma} = 4{,}69 + 3 \cdot 0{,}24 = 5{,}41$

Onderste actiegrens: $\bar{x} - 3\hat{\sigma} = 4{,}69 - 3 \cdot 0{,}24 = 3{,}97$

Voor de spreidingskaart:

target: $\bar{R} = 0{,}27$

Bovenste actiegrens: $D_4 = 3{,}267 \cdot 0{,}27 = 0{,}88$

Onderste actiegrens: : $D_3 = 0 \cdot 0{,}27 = 0$

11.5 Controlekaarten voor attributieve (kwalitatieve) kenmerken

Onder een *attributieve eigenschap* van een product wordt verstaan een eigenschap welke slechts twee waarden kan aannemen; voorbeelden hiervan zijn:
- defect / niet defect;
- te groot / te klein;
- onderdeel aanwezig / afwezig.

Attributieve eigenschappen kunnen worden geteld, resulterend in bijvoorbeeld het aantal fouten k in een steekproef van omvang n. Voor de controle op niveau van attributieve kenmerken kan men gebruik maken van de volgende controlekaarten:
- controlekaart voor *fractie* foutieve exemplaren (zgn. p-kaart);
- controlekaart voor *aantal* foutieve exemplaren (np-kaart);
- controlekaart voor *aantal* fouten (u-kaart).

Voor het bepalen van de controlegrenzen is de voorwaarde dat het type verdeling, met betrekking tot het aantal fouten in een steekproef, bekend is. Een veel voorkomende verdeling in dit verband is de *binomiale verdeling* voor de fractie p en het aantal foutieve exemplaren in een steekproef en de *Poisson-verdeling* voor het aantal *gebeurtenissen.*

Voorbeelden:
Binomiale verdeling: aantal onjuist afgevulde eenheden op een verpakkingslijn in een steekproef van n stuks.
Poisson-verdeling: gemiddeld aantal fouten per spoel vastgesteld in een steekproef van 100 spoelen.

We zullen kort ingaan op de berekening van de controlegrenzen in de verschillende typen controlekaarten.

11.5.1 p-kaart

Voor het bepalen van een p-kaart geldt de volgende voorwaarde.
Het aantal foutieve exemplaren is binomiaal verdeeld. De steekproefomvang is n en p is target van de fractie foutieve exemplaren. Is dit niet het geval, dan dienen de grenzen te worden berekend uit de binomiale verdeling.

a. p-kaart: gegeven norm p

target: p	
Onderste actiegrens:	$p - \sqrt{\dfrac{p(1-p)}{n}}$
Bovenste actiegrens:	$p + \sqrt{\dfrac{p(1-p)}{n}}$

b. p-kaart: geen norm gegeven (p is onbekend)

target: $\bar{p}$	
Onderste actiegrens:	$\bar{p} - 3 \times \sqrt{\dfrac{\bar{p}(1-\bar{p})}{n}}$
Bovenste actiegrens:	$\bar{p} + 3 \times \sqrt{\dfrac{\bar{p}(1-\bar{p})}{n}}$

waarbij: $\bar{p} = \frac{1}{k} \sum_{i=1}^{k} p_i$

p_i = fractie foutieve exemplaren in de i- de steekproef.
k = het aantal steekproeven waaruit $\bar{p}$ wordt berekend.

11.5.2 np-kaart

np is het verwachte aantal foutieve exemplaren in een steekproef van n stuks. De lijnen op de kaart worden berekend door vermenigvuldiging van de lijnen op de p-kaart met n. Stellen we $np = c$, dan krijgen we:

a. np-kaart: *gegeven norm c*

target: c	
Onderste actiegrens:	$c - 3 \times \sqrt{\dfrac{c(n-c)}{n}}$
Bovenste actiegrens:	$c - 3 \times \sqrt{\dfrac{c(n-c)}{n}}$

b. np-kaart: *geen norm gegeven*

target $\bar{c}$	
Onderste actiegrens:	$\bar{c} - 3 \times \sqrt{\dfrac{\bar{c}(n-\bar{c})}{n}}$
Bovenste actiegrens:	$\bar{c} + 3 \times \sqrt{\dfrac{\bar{c}(n-\bar{c})}{n}}$

waarbij: $\bar{c} = \frac{1}{k} \sum_{i=1}^{k} c_i$

c_i = aantal foutieve exemplaren in de i-de steekproef.

11.5.3 u-kaart

u is het aantal fouten per steekproefeenheid. Dit soort controlekaarten kan nuttig zijn als de elementen op de al of niet aanwezigheid van meer dan één eigenschap (in de meeste gevallen door verschillende oorzaken foutief zijn) worden onderzocht.
Voorwaarde:
Het aantal fouten per steekproefeenheid volgt een Poisson-verdeling. Als het gemiddelde aantal fouten $u > \frac{9}{n}$ is, mogen we de grenzen berekenen zoals aangegeven, waarbij n het aantal steekproefeenheden is. Is dit niet het geval dan moeten de grenzen worden berekend uit de Poisson-verdeling.
De lijnen op de kaart worden als volgt berekend:

a. u-kaart: *gegeven norm u*

target: u	
Onderste actiegrens:	$u - 3 \times \sqrt{\frac{u}{n}}$
Bovenste actiegrens:	$u + 3 \times \sqrt{\frac{u}{n}}$

b. u-kaart: *geen norm gegeven*

target: $\bar{u}$	
Onderste actiegrens:	$\bar{u} - 3 \times \sqrt{\frac{\bar{u}}{n}}$
Bovenste actiegrens:	$\bar{u} + 3 \times \sqrt{\frac{\bar{u}}{n}}$

11.6 Testmogelijkheden bij het voeren van controlekaarten

Bij Shewhart-kaarten heeft men naast de grafische weergave van controlekaarten, ook mogelijkheden voor het uitvoeren van statistische tests. Met deze tests is het mogelijk bijzondere afwijkingen te constateren.
Deze tests zijn beschikbaar indien:
- de controlegrenzen de 3σ-grenzen zijn
- de grenzen mogen niet veranderen met de steekproefgrootte

Het doel van deze tests is, specifieke, niet-toevallige afwijkingen in de steekproefmeetwaarden te ontdekken. Het variatiegebied van de meetwaarden is bij deze controlekaart onderverdeeld in gelijke gebieden A, B en C. Wanneer men van de middelste lijn naar boven of beneden gaat komt men achtereenvolgens in gebieden C, B en A. Gebied A ligt dus het verst verwijderd van de middelste lijn. Omdat de gebieden even groot zijn, komen de gebieden C, B en A overeen met 1σ-, 2σ- en 3σ-grenzen.

De statisticus Nelson stelde de volgende testmogelijkheden voor.

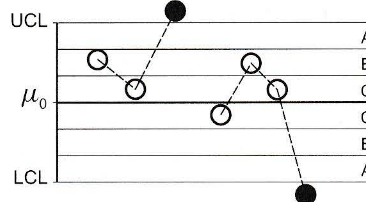
Nummer	Afwijking of aanwezige trend
1	Een punt ligt buiten gebied A, oftewel buiten de 3σ-grenzen, hetgeen duidt op een ontregeld proces, waarbij het niveau en/of de spreiding in het proces is veranderd.
2	Negen punten op rij liggen in gebied C, of daarbuiten, aan één kant van de middelste lijn of normwaarde. Wat duidt op een verschuiving van het procesgemiddelde.
3	Zes punten monotoon stijgend of dalend. Er is een trend in het proces, zodat het proces dreigt te ontsporen. Ingrijpen is noodzakelijk om te voorkomen dat het 'product' moet worden afgekeurd.
4	Veertien punten op rij afwisselen naar boven / naar beneden. Deze zgn. alternerende reeks duidt op een continue wisseling in het proces.
5	Twee van drie punten op rij liggen in gebied A of buiten gebied A. Dit komt door verschuiving van het procesniveau (vergelijkbaar met test 1)
6	Vier van vijf punten op rij liggen in gebied B of buiten gebied B. Het proces heeft een te grote spreiding.
7	Vijftien punten op rij in gebied C (onder en boven middelste lijn). Het proces heeft een veel kleinere spreiding gekregen. De kwaliteit van het proces is beter geworden.
8	Acht punten op rij aan beide zijden van de middelste lijn, maar niet in gebied C. Ook dit duidt op een te grote spreiding in het proces.

Nelson heeft deze toetsingsvoorwaarden zo uitgekozen, dat de mogelijkheid van een puur toevallige 'uitschieter' voor alle tests ongeveer even groot is, namelijk $P < 5\%$, overeenkomend met de onbetrouwbaarheid α bij de toetsingsprocedure.
De tests kan men grafisch weergeven als in figuur 11.5.

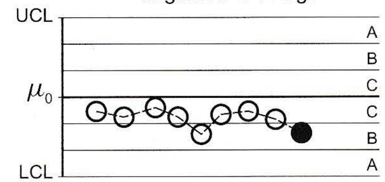
11.6.1 Procescapability-specificatie (C_p en C_{pk})

De 'kwaliteit' van een product of proces wordt vaak weergegeven door *kwaliteitsindices*. In deze kwaliteitindices komt tot uitdrukking de mate van overeenstemming van het product of proces ten opzichte van de opgegeven specificaties van het product of proces. Als een proces statistisch beheerst is, dan is het ook mogelijk om vast te stellen tussen welke grenzen de kenmerken of parameters van een product zullen variëren, wat betreft de kwaliteitseigenschappen. Een gebruikelijke maat voor de productspreiding, van een normaal verdeelde kwaliteitseigenschap, is het '*proces capability interval*'. We weten reeds dat bijna alle meetwaarden van een normaal verdeeld proces zich bevinden tussen $\mu - 3\sigma$ en $\mu + 3\sigma$, waarbij:
μ = procesgemiddelde (= normwaarde van het proces)
σ = processtandaardafwijking (= processpreiding)

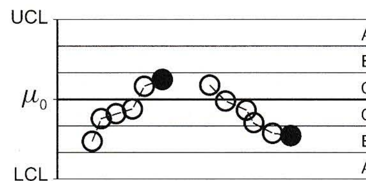
Test 1: Een punt beneden gebied A



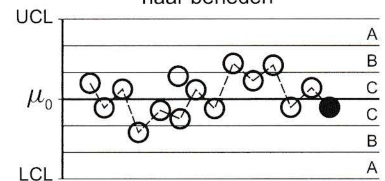
Test 2: Negen punten op een rij in gebied C of lager



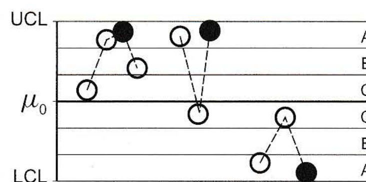
Test 3: Zes punten op een rij monotoon stijgend of dalend



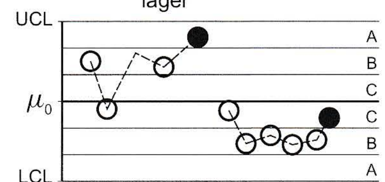
Test 4: Veertien punten op een rij afwisselend naar boven/ naar beneden



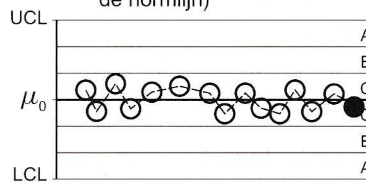
Test 5: Twee van drie punten op een rij in gebied A of lager



Test 6: Vier van vijf punten op een rij in gebied B of lager



Test 7: Vijftien punten op een rij in gebied C (onder en boven de normlijn)



Test 8: Acht punten op een rij aan beide zijden van de normlijn, maar niet in gebied C

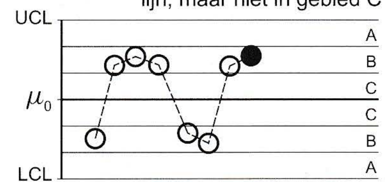


Fig. 11.5

Het procescapability-interval is dan ook gedefinieerd als $[\mu - 3\sigma, \mu + 3\sigma]$. Dit interval zal 99,7% van de individuele producten omvatten.
Het is uiteraard van belang dat de capability van een proces en de productspecificaties op elkaar zijn afgestemd, dit geldt zowel t.a.v. het gemiddelde, als ook t.a.v. de spreiding. Hiertoe zijn verhoudingsgetallen ontwikkeld, die de mate van overeenstemming tot uitdrukking brengen tussen wat men *moet* maken (= de specificatie) en wat men *kan* maken (= proces capability). Hiervoor zijn de laatste jaren de volgende twee *capability indices* naar voren gekomen:

Index voor de spreiding

Bij de afspraken tussen een leverancier en de klant worden specificaties voor bepaalde parameters of kenmerken van het product afgesproken. Meestal wordt dan een onder- en een bovengrens vastgesteld, waarbinnen de waarde van de betreffende parameter zal moeten liggen. Daarnaast hebben we de natuurlijke variaties in het proces, waardoor een bepaalde spreiding in de waarde van de parameter zal ontstaan, wat tot uitdrukking komt in de standaardafwijking σ. Indien de uitkomsten van de parameter een normale verdeling volgen, dan is de totaal mogelijke spreiding (=capability) gelijk aan 6σ. Er is nu een index ontwikkeld ($= C_p$), die de mate van overeenkomst aangeeft tussen de toelaatbare spreiding en de werkelijke spreiding:

$$C_p = \frac{\text{toelaatbare spreiding (= specificatie)}}{\text{werkelijke spreiding (= capability)}} = \frac{USL - LSL}{6\sigma} \tag{11.2}$$

USL = Upper Specification Limit (bovenste specificatiegrens)
LSL = Lower Specification Limit (onderste specificatiegrens)
σ = standaardafwijking van de parameter van het lopend proces

Is de totale spreiding van de parameter ($= 6\sigma$) precies gelijk aan het verschil tussen de onder- en bovengrens van de specificatie, dan is de $C_p = 1$. Elke verandering in het proces en dus verandering in de spreiding van het kenmerk of parameter komt in de C_p- waarde naar voren.

Index voor centrering en spreiding

Naast de spreiding wil men ook vaak iets kunnen zeggen over de centrering van het proces. Dat wil zeggen hoe goed het gemiddelde van de parameter ligt tussen de afgesproken onder- en bovengrens. Om dit vast te leggen heeft men de index voor centrering ($= C_{pk}$) vastgelegd. De C_{pk} is gelijk aan de kleinste waarde van:

$$\frac{\mu - LSL}{3\sigma} \text{ en } \frac{USL - \mu}{3\sigma}$$

μ = gemiddelde van de parameter van het lopend proces
σ = standaardafwijking van de parameter van het lopend proces

Ligt μ precies in het midden tussen de onder- en bovengrens van de afgesproken specificatie en de totale spreiding (6σ) komt precies overeen met het verschil tussen onder- en bovengrens, dan is de $C_{pk} = 1$. Alle verschuivingen van het gemiddelde en/of veranderingen in de spreiding van de parameter komen tot uitdrukking in de C_{pk}- waarde.
Bij C_p- en/of C_{pk}- waarden kleiner dan 1 heeft men een proces dat *niet* voldoet aan de afgesproken specificaties. Als de C_p- en/of C_{pk}- waarden groter zijn dan 1, dan kan het proces in ruime mate voldoen aan de afgesproken specificaties. In de praktijk wordt vaak gesteld dat de C_p- en C_{pk}- waarden groter dan 1,3 moeten zijn. Er kunnen dan kleine verschuivingen in het proces plaatsvinden, zonder dat dit gevolgen heeft voor de kwaliteit van het product (voldoen aan de afgesproken specificaties). In een aantal voorbeelden zullen we het bovenstaande toelichten. Voor de eenvoud nemen we een 'gestandaardiseerd' proces, waarbij de specificaties voor de ondergrens en bovengrens resp. zijn vastgelegd op -3 en +3.

Voorbeeld 3

Als eerste voorbeeld nemen we een proces, waarvan de spreiding van de parameter juist goed is (= gelijk aan de specificatie) en bovendien precies goed is gecentreerd. In figuur 11.6 is de controlekaart van een dergelijk proces weergegeven.

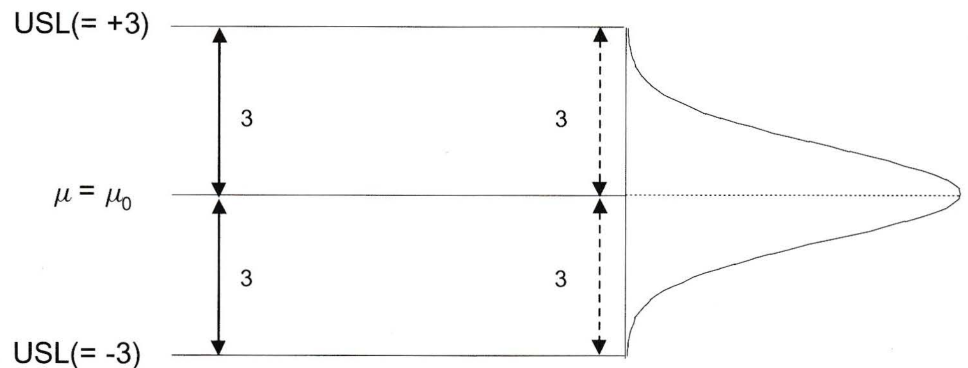


Fig. 11.6 Proces met $C_p = C_{pk} = 1$

De $C_p = \dfrac{USL - LSL}{6\sigma} = \dfrac{+3-(-3)}{6 \times 1} = \dfrac{6}{6} = 1$

Voor de C_{pk} nemen we de kleinste waarde van $\dfrac{USL - \mu}{3\sigma} = \dfrac{+3-0}{3 \cdot 1} = 1$ en $\dfrac{\mu - LSL}{3\sigma} = \dfrac{0-(-3)}{3 \cdot 1} = 1$. Dus $C_{pk} = 1$.

Als het procesgemiddelde en de processpreiding van de specificaties gaat afwijken, dan vindt men dit direct terug in de waarden voor de C_p en de C_{pk}. Een aantal typerende gevallen is in de volgende voorbeelden weergegeven.

Bij het proces van het vorig voorbeeld is de centrering gelijk gebleven, maar de processpreiding is kleiner geworden. ($\sigma_{nieuw} = 0{,}7$). De capability-indices zijn nu respectievelijk:

$C_p = \dfrac{+3-(-3)}{6 \cdot 0{,}7} = 1{,}43$

C_{pk} is kleinste waarde van $\dfrac{+3-0}{3 \cdot 0{,}7} = 1{,}43$ en $\dfrac{0-(-3)}{3 \cdot 0{,}7} = 1{,}43$.

Dus: $C_p = 1{,}43$ en $C_{pk} = 1{,}43$

In figuur 11.7 is de controlekaart weergegeven met de gegevens van dit voorbeeld.

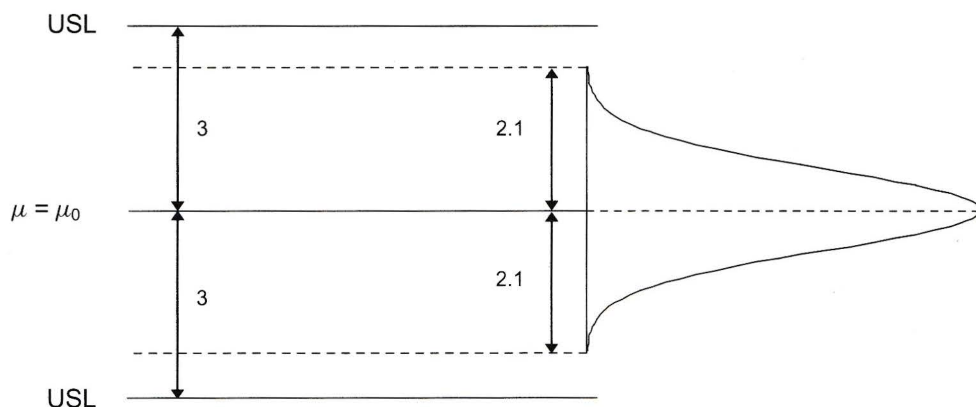


Fig. 11.7 Proces met C_p en $C_{pk} = 1{,}43$

Voorbeeld 4

Stel dat het procesgemiddelde 2 eenheden ten opzichte van de oorspronkelijke waarde naar boven is verschoven, bij gelijkblijvende processpreiding. Dus $\mu = \mu_0 + 2$ en $\sigma = 1$.
De capability-indices zijn nu respectievelijk:

$C_p = \frac{+3-(-3)}{6 \cdot 1} = \frac{6}{6} = 1$

$C_{pk} = \frac{+3-2}{3 \cdot 1} = \frac{1}{3} = 0{,}33$ of $\frac{+2-(-3)}{3 \cdot 1} = \frac{5}{3} = 1{,}67$. Dus $C_{pk} = 0{,}33$ (kleinste van de twee waarden).

Dus: $C_p = 1$ en $C_{pk} = 0{,}33$.

In figuur 11.8 is het proces van dit voorbeeld grafisch weergegeven.

We geven ten slotte een voorbeeld, waarbij de centrering van het proces goed is, maar de standaardafwijking 2 keer zo groot is geworden ($\mu = \mu_0$ en $\sigma = 2$ eenheden).
De capability-indices voor dit proces worden:

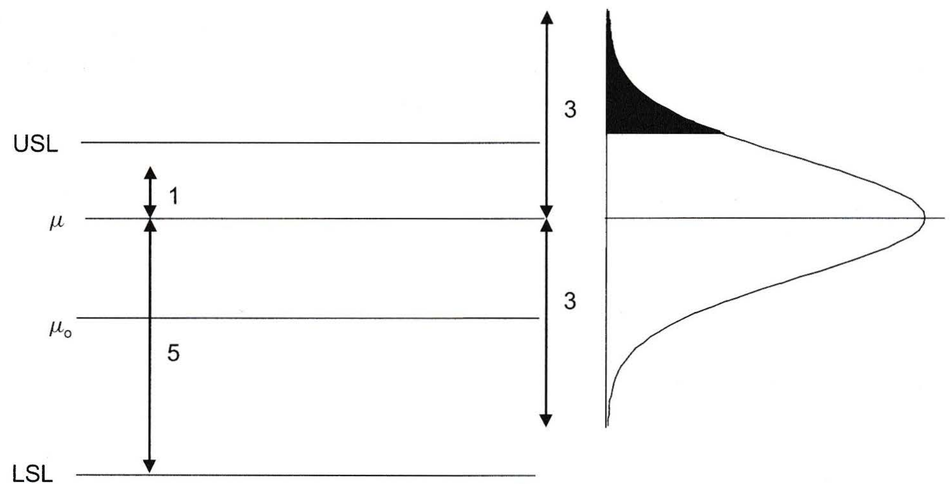


Fig. 11.8 Proces met $C_p = 1$ en $C_{pk} = 0{,}33$

$$C_p = \frac{+3-(-3)}{6\cdot 2} = \frac{6}{12} = 0{,}5$$

$$C_{pk} = \frac{+3-0}{3\cdot 2} = 0{,}5 \text{ of } \frac{0-(-3)}{3\cdot 2} = 0{,}5.$$

Dus: $C_p = C_{pk} = 0{,}5$

We zien dat de centrering goed is, maar doordat de spreiding groter is geworden, is toch de C_{pk} kleiner geworden. De C_{pk}- waarden reageren op veranderingen in niveau en/of spreiding van het proces.

In figuur 11.9 is het proces met $C_p = 0{,}5$ en $C_{pk} = 0{,}5$ weergegeven.

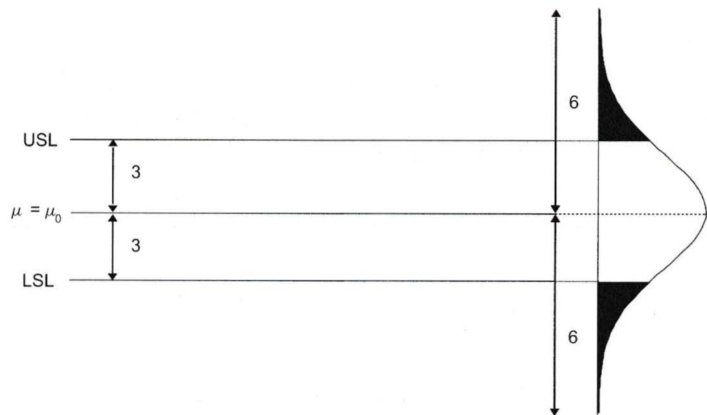


Fig. 11.9 Proces met $C_p = 0{,}5$ en $C_{pk} = 0{,}5$

Opgaven

1. Voor een procescontrole wordt regelmatig een smeltpuntsbepaling uitgevoerd. Bij een normaal verlopend proces is het gemiddelde smeltpunt $m = 89$ °C, met standaardafwijking $s = 1{,}5$ °C.
De smeltpunten volgen hierbij een normale verdeling.
Stel van dit proces een controlekaart op voor het gemiddelde van 3 metingen.

2. Teneinde de viscositeit van een grondstof te controleren wordt dagelijks uit de productie een steekproef van vier monsters genomen. Van 20 dagen zijn de 20×4 uitkomsten in onderstaande tabel vermeld (in seconden).

1	2	3	4	5	6	7	8	9	10
324	328	346	333	329	323	309	334	322	300
331	329	315	307	315	311	313	313	301	336
317	333	304	326	317	307	321	313	312	316
330	320	300	324	310	313	321	318	322	314

11	12	13	14	15	16	17	18	19	20
314	318	336	323	319	313	319	324	312	320
321	319	315	317	317	311	313	312	316	321
327	323	314	326	321	312	321	324	328	319
318	326	330	331	327	330	318	325	315	319

Maak van deze gegevens een controlekaart voor steekproefgemiddelden en standaardafwijkingen van 4 waarnemingen ($\bar{x}$-s-kaart).

3. We gaan er nu van uit dat de viscositeitscontrole van bovengenoemde grondstof (zie opgave 2) wordt gedaan, door per dag slechts één monster te nemen. Hierbij zijn de volgende uitkomsten verkregen:

1	2	3	4	5	6	7	8	9	10
324	328	346	333	329	323	309	334	322	300

11	12	13	14	15	16	17	18	19	20
314	318	336	323	319	313	319	324	312	320

Maak van bovenstaande gegevens een $\bar{x} - R$-kaart.

4. Een product wordt afgeleverd in zakken van 50 kg. Om het afvulproces te controleren, wil men een controlekaart gaan inrichten. Gedurende een periode van 20 dagen, waarin het afvulproces redelijk beheerst is, neemt men per dag een steekproef van 3 zakken. Van elke zak wordt het gewicht bepaald. De volgende resultaten zijn verkregen:

dag	meetuitkomsten (in kg)		
1	50,3	50,5	49,6
2	50,1	50,0	49,5
3	50,3	50,1	50,2
4	50,1	50,3	50,2
5	50,3	49,8	49,7
6	50,2	50,1	50,0
7	49,7	50,2	50,3
8	50,0	50,0	50,5
9	49,5	49,9	50,0
10	50,3	50,1	50,4
11	50,0	50,0	49,6
12	50,7	50,6	50,3
13	50,4	50,0	50,1
14	49,7	49,9	50,3
15	49,9	49,7	50,1
16	50,1	50,4	50,3
17	49,7	49,5	50,1
18	49,8	49,3	49,9
19	50,2	50,4	50,1
20	49,7	49,9	50,4

Maak met behulp van bovenstaande gegevens een $\bar{x}$-s- kaart voor steekproeven van $n = 4$.

5. Bij de serieproductie van een artikel wordt regelmatig een exemplaar gecontroleerd. Om nu te komen tot een controlekaart voor deze productie, neemt men gedurende 20 weken steekproeven van ca 200 exemplaren per week. In elke steekproef wordt het aantal exemplaren, dat niet aan de specificaties voldoet, genoteerd. De volgende gegevens zijn hierbij verkregen:

week	aantal onderzochte exemplaren	aantal foutieven	fractie foutieven
1	210	12	0,0571
2	198	21	0,1060
3	196	18	0,0918
4	210	23	0,1095
5	190	7	0,0368
6	200	18	0,0900
7	210	13	0,0619
8	220	25	0,1136
9	200	12	0,0600
10	218	15	0,0688
11	206	19	0,0922
12	196	18	0,0918
13	190	23	0,1210
14	196	9	0,0459
15	198	13	0,0656
16	206	14	0,0679
17	210	17	0,0809
18	204	20	0,0980
19	196	21	0,1071
20	200	9	0,0450
totaal	4054	327	1,6115

Construeer, met behulp van bovenstaande gegevens een p-kaart. (Voor n neemt men de gemiddelde steekproefgrootte).

6. Construeer met behulp van de gegevens van opgave 5 een np-kaart.

7. Een product wordt samengesteld uit verschillende componenten. Bij dit samenstellen kunnen meerdere fouten optreden. De productieleiding wil nu een controlekaart voor het gemiddeld aantal fouten. Uit de productie neemt men daartoe over langere tijd 20 steekproeven van elk 10 exemplaren. Per steekproef telt men het aantal fouten en berekent vervolgens het gemiddeld aantal fouten per steekproef. De uitkomsten staan in onderstaande tabel:

steekproef	aantal fouten	gemiddeld aantal fouten per steekproef
1	17	1,7
2	14	1,4
3	6	0,6
4	23	2,3
5	7	0,7
6	11	1,1
7	4	0,4
8	13	1,3
9	18	1,8
10	21	2,1
11	5	0,5
12	14	1,4
13	8	0,8
14	9	0,9
15	15	1,5
16	7	0,7
17	3	0,3
18	19	1,9
19	17	1,7
20	21	2,1
$\sum$	252	25,2

Construeer een u-kaart met behulp van bovenstaande gegevens

8. Met een leverancier zijn de volgende tolerantiegrenzen afgesproken:
Bovengrens $(USL) = 2{,}650$ mg en ondergrens $(LSL) = 2{,}350$ mg.
Om te controleren of het proces hieraan voldoet, neemt men 20 steekproeven van $n = 4$.
De berekende procesparameters hieruit zijn:
- procesgemiddelde $\mu = 2{,}500$ mg
- processpreiding $\sigma = 44{,}2$ mg

Bereken de C_p- en de C_{pk}-waarden. Is de centrering en/of de spreiding van het proces goed?

9. De procesparameters van een proces, bij 20 steekproeven van $n = 4$ zijn:
- procesgemiddelde $\mu = 255$
- gemiddelde spreidingsbreedte $\bar{R} = 11$

Als de C_{pk} waarde 1,2 moet bedragen, wat moeten de gespecificeerde tolerantiegrenzen zijn bij een precies goed gecentreerd proces?

10. Van een proces zijn de volgende gegevens bekend:
Target: $\mu = 32{,}5$
Spreiding: $\sigma = 2{,}7$
Het proces wordt gecontroleerd door steekproeven te nemen van $n = 5$.
De gespecificeerde grenzen bedragen: $USL = 35{,}6$ en $LSL = 28{,}5$.

- Bereken de C_p- en de C_{pk}-waarden.
- Hoeveel procent van de productie valt buiten de grenzen?

Bijlage A Statistiek met EXCEL

Microsoft EXCEL beschikt over tal van statistische functies. Hoewel EXCEL niet in eerste instantie ontworpen is als een softwarepakket voor statitistische programmatuur (zoals bij SPSS, SAS, Minitab wel het geval is) biedt het ruim voldoende toepassingsmogelijkheden voor hoger opgeleiden die niet als gespecialiseerd statisticus zullen werken, maar slechts af en toe statistiek zullen gebruiken. Twee grote voordelen van EXCEL ten opzichte van statistische programmatuur zijn dat het gemakkelijk (en goedkoop) beschikbaar is en eenvoudig. Vrijwel alle in dit boek genoemde formules en tabellen kunnen op eenvoudige wijze gebruikt worden. In deze bijlage geven we een opsomming van de mogelijkheden van EXCEL ten behoeve van de technieken in de verschillende hoofdstukken van dit boek. In de meeste gevallen zullen we een voorbeeld geven als illustratie. Een aantal voorbeelden zijn uit de hoofdstukken van dit boek afkomstig.

A.1 Inleiding

We gaan ervan uit dat de lezer enigzins bekend is met het werken in spreadsheets. Is dit niet het geval dan raden we aan dat de lezer zich eerst vertrouwd maakt met de basishandelingen, zoals het werken met de muis, het gebruik van de werkbalken, het invoeren van gegevens in cellen (en het daarbij behorende format), het verschil tussen relatieve en absolute celadressen, het kopiëren en plakken van kolommen met gegevens en functies, het opmaken van een worksheet enzovoorts. In deze bijlage worden deze handelingen soms gedetailleerd beschreven, maar zeker niet altijd.

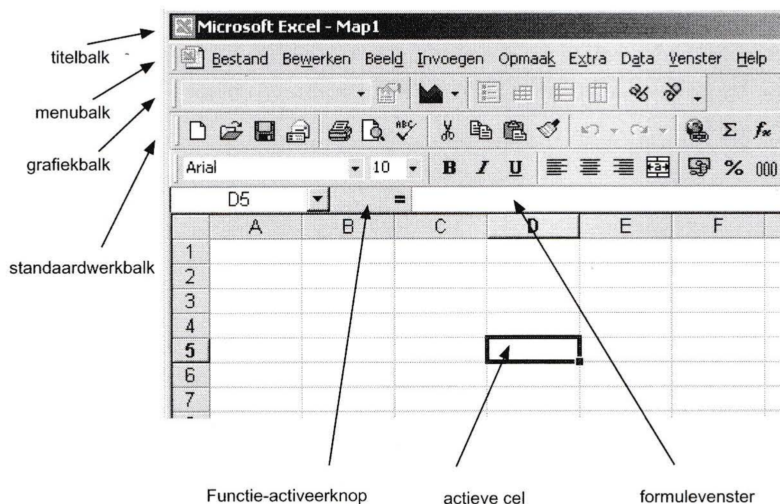


Fig. A1 Gedeelte van het bureaublad

Het moge bekend zijn dat EXCEL een zeer gebruiksvriendelijke help-functie heeft. Er zijn verschillende versies van EXCEL in omloop. De versies EXCEL 97 en EXCEL 2000 verschillen nauwelijks. Voor de in deze bijlage genoemde toepassingen zijn er zelfs helemaal geen verschillen tussen deze laatste twee versies. In figuur A1 zien we een gedeelte van het bureaublad met de belangrijkste begrippen. We kunnen het bureaublad zelf aanpassen onder **Beeld** en dan **Werkbalken** (aanvinken welke van toepassing zijn).

EXCEL heeft standaard een groot aantal statistische functies, waarvan we er een aantal zullen laten zien. Bovendien beschikt EXCEL over een groep functies onder de noemer 'Gegevensanalyse', die meestal nog niet direct bij de installatie zijn inbegrepen. Om deze te activeren, klik op de werkbalk bij **Extra** en vervolgens bij **Invoegtoepassingen**. Vink dan in elk geval 'Analysis ToolPak' en 'Analysis ToolPak-VBA' aan en activeer met OK. In het menu **Extra** is nu **Gegevensanalyse** toegevoegd (zie figuur A2).

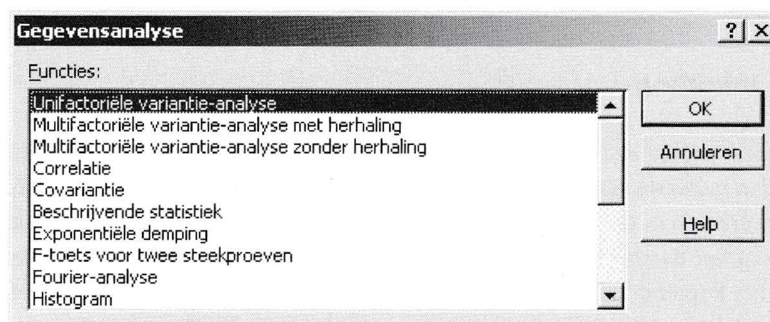


Fig. A2 Extra beschikbare functies voor gegevensanalyse

Van deze nieuwe mogelijkheden zullen we er een aantal bekijken. De functies die in EXCEL reeds aanwezig zijn, kunnen we zichtbaar maken door bijvoorbeeld op het =-teken naast het formulevenster te klikken.
Het resultaat is:

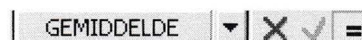


Fig. A3

op de formulebalk, met links de laatst gebruikte functie (hier: GEMIDDELDE) waarna we onder het pijltje een overzicht krijgen van alle beschikbare functies. Onder 'Meer functies' kunnen we het overzicht krijgen van alle statistische functies, waarover EXCEL beschikt.

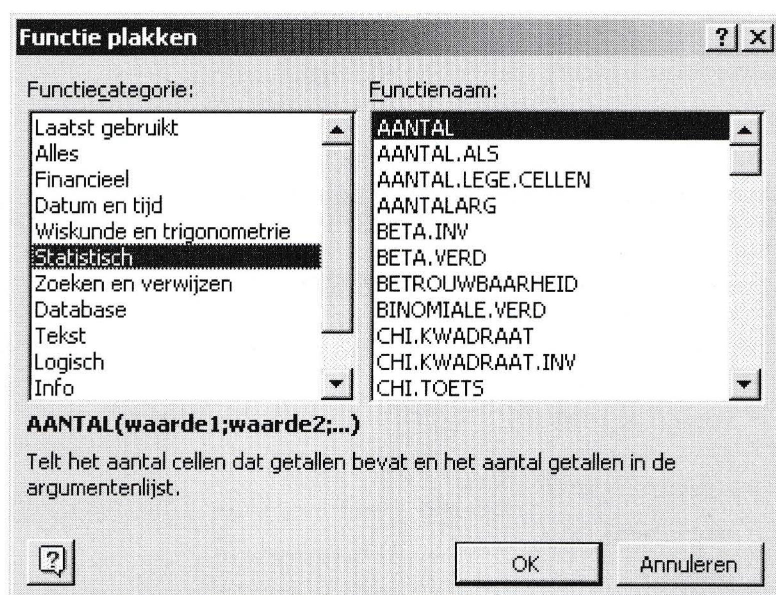


Fig. A4

De functie die we nu aanklikken, verschijnt in het formulevenster. Een andere manier om functies te activeren is te klikken op de f_x-knop in de werkbalk (naast het $\sum$-teken).

A.2 Beschrijvende statistiek

In de hoofdstukken 1, 2 en 3 van dit boek hebben we de beschrijvende statistiek behandeld. We zullen nu aan de hand van voorbeeld 1 uit hoofdstuk 3 zien hoe we een steekproef of populatie kunnen beschrijven.

In voorbeeld 1 hadden we 50 afgeronde gewichten. Voorzien van een label ('gewicht') plaatsen we deze gewichten in een kolom. Voor een beter overzicht kunnen we de 50 getallen beter sorteren (bijvoorbeeld bij **Data**, **Sorteren**). Een gedeelte (de eerste 7 gesorteerde waarnemingsuitkomsten) ziet er ongeveer zo uit:

	A
1	gewicht
2	81
3	84
4	86
5	87
6	87
7	87
8	88

Fig. A5

We kunnen het snelst de 50 waarnemingsuitkomsten analyseren door onder **Extra** en **Gegevensanalyse** 'Beschrijvende statistiek' te activeren.

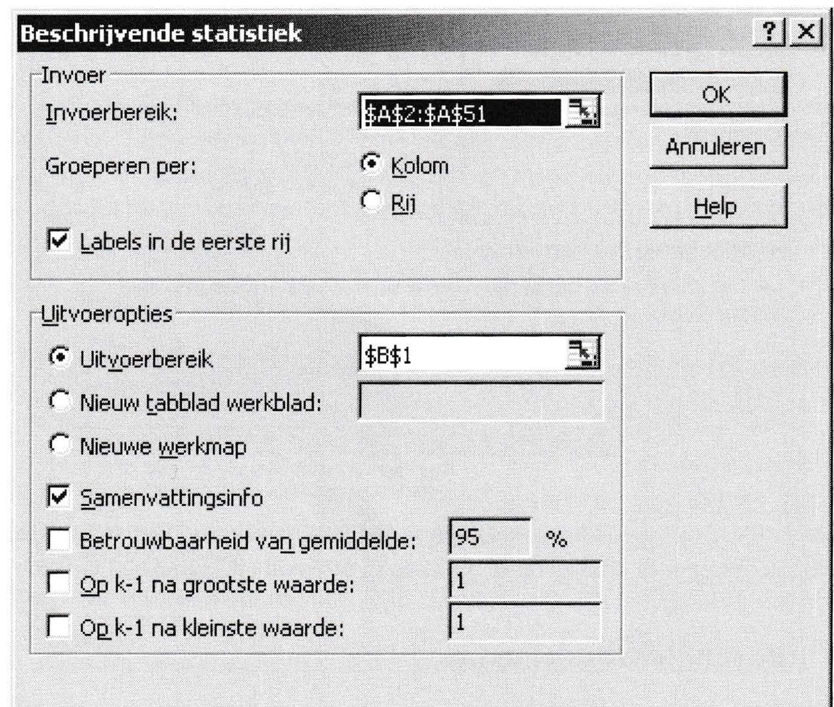


Fig. A6

Voor het 'Invoerbereik' markeren we met de cursor de gehele A-kolom t/m de laatste waarnemingsuitkomst. Denk erom 'Labels in de eerste rij' aan te vinken (op de eerste rij staat een label, geen waarnemingsuitkomst). Onder uitvoerbereik selecteren we met de muis bijvoorbeeld cel B1. Vink ook 'Samenvattingsinfo' aan, zodat het dialoogvenster er als hierboven ziet.

Na het activeren (OK) ontvangen we de samenvattingsinformatie van alle waarnemingsuitkomsten (maak eventueel kolom B en kolom C breder om het resultaat beter te kunnen zien):

B	C
gewicht	
Gemiddelde	94,1
Standaardfout	0,842857143
Mediaan	94,5
Modus	96
Standaarddeviatie	5,959900013
Steekproefvariantie	35,52040816
Kurtosis	1,193717868
Scheefheid	0,62890393
Bereik	30
Minimum	81
Maximum	111
Som	4705
Aantal	50

Fig. A7

Kurtosis en scheefheid zijn niet behandeld in dit boek. Maar uit de rest van de informatie kunnen we veel opmaken. Nu nog de grafiek. Dit kunnen we EXCEL direct en geheel zelf laten doen onder **Extra**, **Gegevensanalyse**, 'Histogram'. Onder 'Invoerbereik' selecteren we weer kolom A1 t/m A51, klik Labels wederom aan, klik ook 'Grafiek maken' aan en laat de uitvoer (Uitvoerbereik) beginnen bij cel D2 (klik de cel aan). Maak het 'Verzamelbereik' leeg. Na uitvoeren (OK) krijgen we een niet al te fraaie frequentieverdeling, met bijbehorend histogram (we laten alleen de frequentieverdeling zien):

Verzamelbereik	*Frequentie*
81	1
85,28571429	1
89,57142857	10
93,85714286	8
98,14285714	21
102,4285714	4
106,7142857	3
Meer	2

Fig. A8

We kunnen ook zelf een (mooiere) klassenindeling maken en wel als volgt.

Uit de verkregen gegevens kunnen we opmaken dat de Range 30 is (111-81). Volgens de formule

$$b = \frac{R}{\sqrt{n}} = \frac{30}{\sqrt{50}} \approx 4{,}2$$

kunnen we besluiten even brede klassen te maken met breedte 5 (4,2 afgerond naar boven). Er zijn dan 7 klassen nodig. In EXCEL moeten we nu de *bovengrenzen* van deze klassen in een kolom zetten. We kiezen analoog aan het voorbeeld in hoofdstuk 3 voor de bovengrenzen 84,5 - 89,5 - 94,5 - enzovoorts. Doe dit als volgt: plaats de label 'bovengrens' bijvoorbeeld in cel B36, plaats daaronder 84,5, daaronder 89,5, markeer de laatste twee cellen (B37 en B38), ga met de cursor in het gemarkeerde gebied staan totdat het plusteken verschijnt en sleep naar beneden totdat alle bovengrenzen gemaakt zijn. Op een schone plaats bijvoorbeeld vanaf cel B27 kunnen desgewenst de ondergrenzen geplaatst worden. Neem aan dat de 7 bovengrenzen (inclusief label bovengrens) staan in de cellen B36 t/m B43. Ga nu naar **Extra**, **Gegevensanalyse**, Histogram en voer als volgt in:

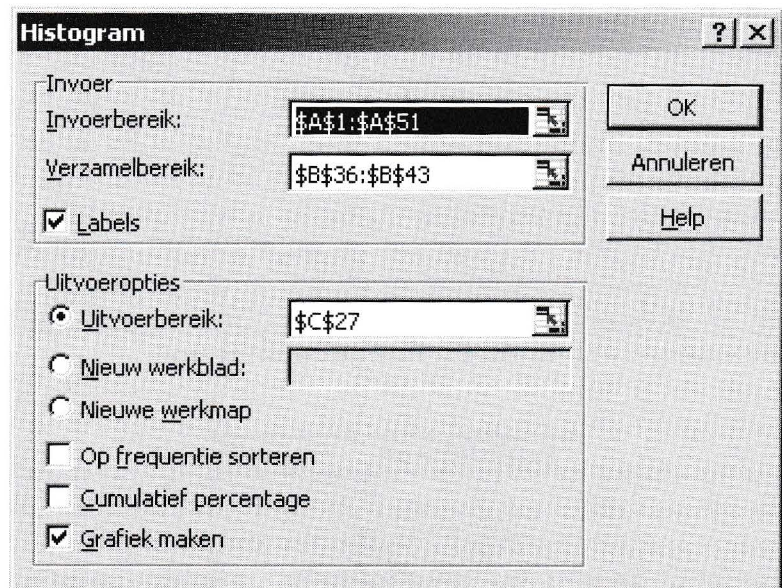


Fig. A9

Na OK verschijnt de frequentieverdeling zoals we die willen hebben, plus een histogram:

ondergrens klasse	bovengrens klasse	Frequentie
79,5	84,5	2
84,5	89,5	10
89,5	94,5	13
94,5	99,5	20
99,5	104,5	2
104,5	109,5	1
109,5	114,5	2
	Meer	0

Fig. A10

Het histogram ziet er misschien niet zo fraai uit (te lage staven bijvoorbeeld). Door met de cursor op diverse plaatsen in het histogram te gaan staan, kan het formaat vergroot of verkleind worden door de rand te verslepen. Het resultaat zou er bijvoorbeeld zo uit kunnen zien:

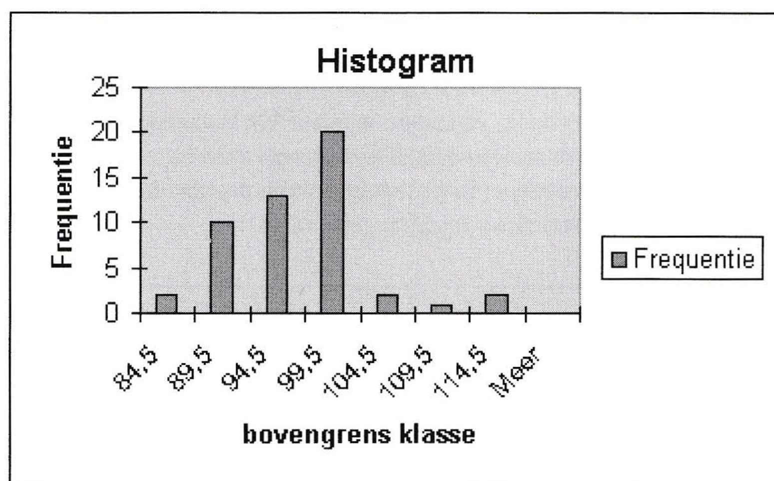


Fig. A11

Er kan van alles aan het histogram veranderd worden. Steeds door het gedeelte dat we willen veranderen te selecteren met de muis, vervolgens met de rechtermuisknop klikken en de veranderingen (titel, legenda, kleuren, enzovoorts) aan te brengen. In onderstaande figuur is bijvoorbeeld de *breedte* van de staven (klik op de rechter-of linkerrand (niet in het midden!) van een staaf totdat deze *rondom* gemarkeerd wordt, rechtermuisknop indrukken, Gegevenspunt opmaken, Opties, Breedte tussenruimte op nul stellen) aangepast en de *titel*

(Klik op titel Histogram in vorige grafiek, totdat de titel rondom gemarkeerd wordt, waarna de tekst te veranderen is) veranderd. We raden de lezer aan met het aanpassen van het histogram zelf te experimenteren.

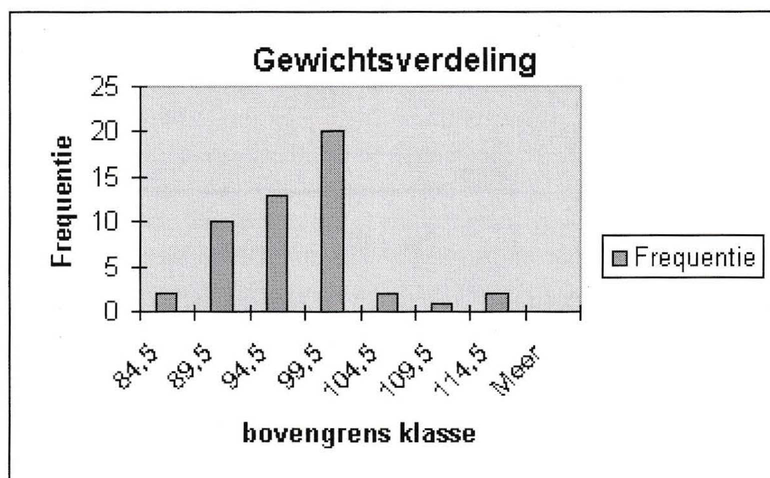


Fig. A12

Ook de soort grafiek kan eenvoudig veranderd worden. We kunnen kiezen uit vele soorten, bijvoorbeeld de polygoon. We moeten de klassenbovengrenzen dan wel vervangen door de klassenmiddens. Op de categorie-as (horizontale as) zijn daartoe de maatstreepjes aangepast. In het tekengebied zijn ook rasterlijnen ingevoerd.

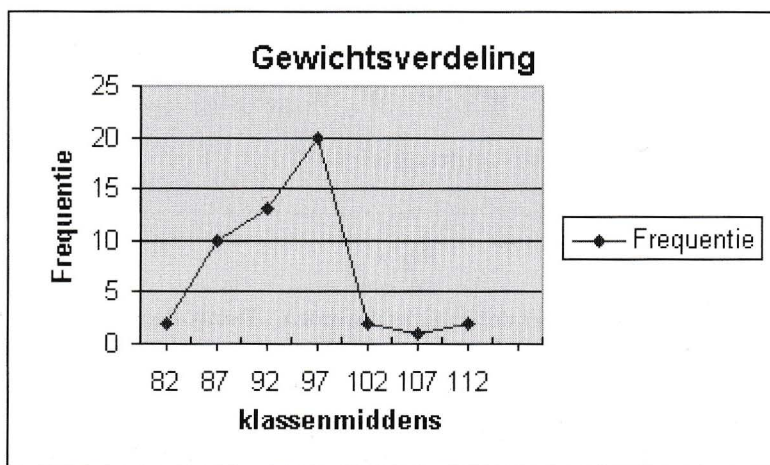


Fig. A13

Tot zover de meeste gebruikte functies van de beschrijvende statistiek. Overige functies die in dit kader eventueel gebruikt kunnen worden, zijn GEMIDDELDE, STDEV (standaardafwijking steekproef), STDEVP (standaardafwijking populatie), MEDIAAN, MODUS, KWARTIEL en PERCENTIEL.

A.3 Kansberekeningen

In hoofdstuk 4 van het boek werden de kansregels uiteengezet. Bij het berekenen van kansen kon soms handig gebruikgemaakt worden van de formules voor permutaties, combinaties en variaties. De functie PERMUTATIES zit onder 'Statistische functies'. De formule voor COMBINATIES zit in EXCEL onder 'Wiskundige en trigonometrische functies'.
Met PERMUTATIES(N;M) wordt het aantal manieren berekend waarop M uit N objecten kunnen worden gepermuteerd. Hierbij tellen alle mogelijke volgorden mee.
Met COMBINATIES(M;N) wordt het aantal rangschikkingen van M uit N berekend zonder de volgorde in acht te nemen. COMBINATIES(15;2) geeft als antwoord 105 en PERMUTATIES(15;2) geeft als antwoord 210. Dit is tevens het aantal *variaties* van 2 uit 15.

A.3.1 Discrete kansverdelingen

We kunnen in EXCEL zowel zelfgeformuleerde discrete kansverdelingen invoeren en doorrekenen als de bekende kansverdelingen zoals de binomiale verdeling, de hypergeometrische verdeling en de Poisson-verdeling.
We geven eerst een voorbeeld van een zelf ingevoerde verdeling en nemen daarvoor het voorbeeld uit paragraaf 5.2, betreffende de som van de ogenaantallen bij een worp met twee dobbelstenen. Stel dat we de kans willen berekenen dat de som tussen 6 en 8 ligt (grenzen meegerekend), dus:

$$P(6 \leq K \leq 8)$$

De betreffende kansverdeling voeren we in twee kolommen in. De functie KANS vraagt om de volgende gegevens (de breuken in de tweede kolom blijven breuken door de celeigenschappen zodanig in te stellen: markeer de cellen, rechts klikken en Celeigenschappen instellen op 'breuken'):

	A	B
1	k	kans
2	2	1/36
3	3	1/18
4	4	1/12
5	5	1/9
6	6	5/36
7	7	1/6
8	8	5/36
9	9	1/9
10	10	1/12
11	11	1/18
12	12	1/36

Fig. A14

x-bereik: hiervoor moeten de waarden 2 t/m 12 met de cursur gemarkeerd worden, de bijbehorende cellen (A2:A12) verschijnen in het venster. Om te kunnen markeren, kunnen we het beste het dialoogschermpje van KANS even verstoppen door te klikken op het knopje met het pijltje erin rechts in het venster van het x-bereik). Na OK komt het dialoogscherm weer terug).
kansbereik: hiervoor moeten de bijbehorende kansen gemarkeerd (B2:B12) worden.
ondergrens: voer 6 in.
bovengrens: voer 8 in.
Na activering (OK) verschijnt het antwoord (hier 0,4444) in de cel waar de cursor op dat moment staat.
Wanneer we de verwachtingswaarde en de variantie (of de standaardafwijking) van een discrete kansvariabele willen bepalen, is er wat meer EXCEL-kennis nodig. We zullen dit laten zien aan de hand van het voorbeeld van de som van de ogenaantallen van de twee dobbelstenen.
We gebruiken de formules $\mu = \sum k_i P(K = k_i)$ en $\sigma^2 = \sum (k_i - \mu)^2 P(K = k_i)$ uit hoofdstuk 5 en gaan daarmee (uitgaande van de gegevens die zojuist in de A- en de B-kolom zijn ingevoerd) als volgt te werk.

- Geef de tekst 'gemiddelde' in cel C1 en de tekst 'variantie' in cel D1.
- Selecteer cel C2 en typ een =-teken (EXCEL weet nu dat er een formule aankomt).
- Ga naar cel A2, klik met de linkermuisknop (er verschijnt een *relatief* adres na het =-teken, want er zit geen $-teken in), plaats een maal-teken (*), verplaats de cursor naar cel B2 en klik wederom links. In cel C2 hoort nu de formule =A2*B2 te staan. Met Enter wordt dit product berekend.
- Kopieer nu de formule in cel C2 naar de rest van kolom C. De makkelijkste manier is: ga op de rand van cel C2 staan tot het +-teken (het dunne, niet het dikke plusteken!) verschijnt en sleep nu naar beneden met de linkermuisknop (we bereiken hetzelfde onder **Doorvoeren** in het menu **Bewerken**).
- Ten slotte worden alle producten in kolom C opgeteld door op het $\sum$-teken te klikken, staande in cel C13, gevolgd door Enter.

Het resultaat moet 7 zijn, zoals verwacht.
De variantie berekenen we als volgt.

- Ga naar cel D2 en typ een =-teken, gevolgd door een linker haakje (..
- Ga naar cel B1, klik links, typ een minteken.
- Nu moet μ ingevuld worden. Deze waarde staat in cel C13. Ga dus naar cel C13. Wanneer we cel C13 direct zouden kopiëren, zou een relatieve waarde verschijnen: deze verandert mee als we gaan kopiëren. Er moet een *absolute* waarde (constant) verschijnen. De makkelijkste manier om dit te bereiken, is met de functietoets F4. Na het minteken in cel B1 komt nu C13 te staan.
- Sluit nu af met een haakje, kwadrateer (^2), plaats een maal-teken en klik links op cel B2. Sluit af met Enter.
- Kopieer nu cel D2 naar de overige cellen in kolom D en sommeer.

Het resultaat moet er ongeveer zo uitzien (merk op dat alle cellen breuken bevatten, in het algemeen hoeft dit natuurlijk niet):

	A	B	C	D
1	k	kans	gemiddelde	variantie
2	2	1/36	1/18	25/36
3	3	1/18	1/6	8/9
4	4	1/12	1/3	3/4
5	5	1/9	5/9	4/9
6	6	5/36	5/6	5/36
7	7	1/6	1 1/6	0
8	8	5/36	1 1/9	5/36
9	9	1/9	1	4/9
10	10	1/12	5/6	3/4
11	11	1/18	11/18	8/9
12	12	1/36	1/3	25/36
13			7	5 5/6

Fig. A15

A.3.2 De binomiale verdeling

Voor het werken met de binomiale verdeling beschikt EXCEL over de functie BINOMIALE.VERD.

BINOMIALE.VERD

Aantal-gunstig	2	= 2
Experimenten	20	= 20
Kans-gunstig	0,3	= 0,3
Cumulatief	2	= WAAR

= 0,035483132

Geeft als resultaat de binomiale verdeling.

Aantal-gunstig is het aantal gunstige uitkomsten in een experiment.

? Resultaat formule = 0,035483132 OK Annuleren

Fig. A16

De aanroep van deze functie lichten we toe aan de hand van voorbeeld 16 uit het boek.

Het ging hierbij om de vraag hoe groot de kans is dat iemand minstens drie prijzen wint, wanneer hij 20 loten koopt. Voor elk lot geldt dat de kans op een prijs 0,3 bedraagt. We berekenen daartoe de kans dat iemand hoogstens twee 'successen' boekt en passen hierop de complementregel toe. Zoals uit figuur A16 blijkt, moet worden ingevuld:

Aantal-gunstig = het aantal successen (=2)

Experimenten = de steekproefgrootte (=20)

Kans-gunstig = de fractie (0,3, geen decimale punt maar een komma, althans in de nederlandse versie van EXCEL). Bij *cumulatief* vullen we WAAR in, immers we berekenen $P(K \leq 2)$, dit is een cumulatieve kans.

Het antwoord verschijnt ook op de plaats waar de cursor staat (0,035483132), zodat de gevraagde kans volgens de complementregel ongeveer 0,9646 bedraagt.

A.3.3 De hypergeometrische verdeling

Bij steekproeven zonder teruglegging uit kleine populaties (niet-constante fractie) is het aantal successen hypergeometrisch verdeeld. EXCEL heeft hiervoor de beschikking over de functie HYPERGEO.VERD. Als voorbeeld van een toepassing hiervan nemen we voorbeeld 20 uit hoofdstuk 5.

In een partij van 40 computers zitten 3 defecte. Wanneer men uit deze partij zonder teruglegging 2 computers neemt, wat is dan de kansverdeling van het aantal defecte computers in de steekproef? Het kenmerk waar in de steekproef op wordt gelet is dus 'defect'.

Bij aanroep van de functie HYPERGEO.VERD wordt gevraagd:

Steekproef-gunstig = aantal defecte computers in de steekproef (hier 0, 1 of 2).

Grootte-steekproef = 2 (spreekt voor zichzelf).

Populatie-gunstig = het aantal defecte computers in de populatie (=4).

Grootte-populatie = 40 (spreekt voor zichzelf).

We rekenen de gevraagde kans voor $K = 0$ uit, voor andere waarden van K gaat het net zo. Het antwoord verschijnt ook in de cel waar de cursor op dat moment in de spreadsheet staat.

HYPERGEO.VERD

Steekproef-gunstig	0	= 0
Grootte-steekproef	2	= 2
Populatie-gunstig	4	= 4
Grootte-populatie	40	= 40

= 0,807692308

Geeft als resultaat de hypergeometrische verdeling.

Grootte-populatie Is de grootte van de populatie.

Resultaat formule = 0,807692308 OK Annuleren

Fig. A17

A.3.4 De Poisson-verdeling

Voor de Poisson-verdeling beschikt EXCEL over de functie POISSON. We zullen laten zien hoe het gemakkelijkst een Poisson-verdeling gegenereerd kan worden. We nemen daarvoor voorbeeld 24 uit hoofdstuk 5. Het Poisson-verdeelde kenmerk is hier het aantal telefoongesprekken dat per uur doorkomt. Gegeven is dat dit gemiddeld 4 bedraagt, waarmee de verdeling vastligt. Hoe de verdeling eruitziet, bepalen we als volgt.

- Geef in cel A1 de tekst k en in cel B1 de tekst $P(K = k)$.
- Typ in cel A2 het getal 0 en in cel A3 het getal 1.
- Vul kolom A nu met oplopende waarden, bijvoorbeeld door cel A1 en A2 gezamenlijk te markeren en door het slepen van het +-teken (dat verschijnt op de rand van de gemarkeerde cellen) de getallen 2, 3, 4,...., 15 te genereren.
- Plaats de cursor in cel B2 en roep de POISSON-functie aan. Bij de waarde van X markeren we de cellen A2 t/m A17 in, voor het Gemiddelde geven we het getal 4 (gegeven) in en vul bij Cumulatief ONWAAR in. Na OK verschijnt in cel B2 het getal 0,018316.
- Vanaf cel B2 kunnen we nu met het plusje naar beneden slepen en zien de Poissonkansen verschijnen. Controleer zelf of de som vrijwel 1 is (we zouden nog even door kunnen gaan vanaf $k = 16$ om te bereiken dat de som van alle kansen daadwerkelijk 1 is).

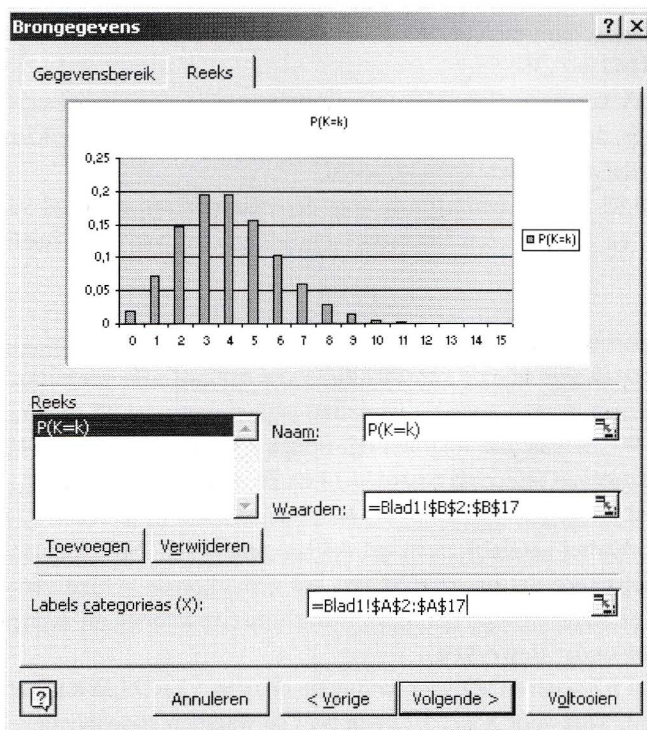


Fig. A18

- Ga nu ergens op een open cel staan en activeer de Grafiek-wizard (gekleurde histogram-ikoon) in de werkbalk. Selecteer Kolom en klik naar het volgende scherm.
- Klik op Reeks en vervolgens Toevoegen.
- In het venster Naam kan $P(K = k)$ vermeld worden.
- In het venster Waarden komt de kolom B2:B17 (selecteren)
- In het venster Labels categorie x-as komt de kolom A2:A17 (selecteren). Het scherm Brongegevens ziet er nu uit als in onderstaande figuur.
- Daarna kan de Wizard voltooid worden.

A.3.5 De normale verdeling

Van de continue kansverdelingen is natuurlijk vooral de normale verdeling van belang. Voor een algemene normale verdeling zijn twee functies te gebruiken.
Met NORM.VERD kan direct de kans berekend worden dat een normaal verdeelde (gemiddelde μ en standaardafwijking σ) variabele X een waarde heeft kleiner dan x. Een cumulatieve kans dus.
Ingevoerd moeten worden (volg voorbeeld 7 uit hoofdstuk 6):
X = de waarde van X ($x = 176{,}4$)
Gemiddelde (= 174)
Standaardafwijking (= 7)
Cumulatief (= WAAR)
Het antwoord verschijnt direct: $P(X < 176{,}4) = 0{,}634147$. De kans dat $P(X > 176{,}4)$ is dus $1 - 0{,}634147 = 0{,}365\,85$.
Met NORMINV berekenen we bij een gegeven linkeroverschrijdingskans (cumulatieve kans $P(X < x) = \alpha$, dus), de bijbehorende x-waarde. Daartoe moeten de kans, het gemiddelde en de standaardafwijking worden ingevoerd.
NORM.INV(0,85; 50; 5) berekent de waarde waarvoor een normaal verdeelde variabele met $\mu = 50$ en $\sigma = 5$ een linkeroverschrijdingskans van 0,85 heeft (antwoord $x = 55{,}18216439$).

De grafiek van een normale verdeling

We zullen nu laten zien hoe de kansdichtheid van een normale verdeling in beeld gebracht kan worden. Laten we een normale verdeling nemen met $\mu = 20$ en $\sigma = 2$. Uit de theorie weten we dat 99,7% van alle waarden ligt tussen $\mu - 3\sigma$ en $\mu + 3\sigma$. We gaan daarom de kansdichtheid tekenen tussen de waarden 14 en 26. Ga als volgt te werk.

- Plaats in cel A1 de tekst 'x' en in cel B1 de tekst '$f(x)$'.
- Typ in cel A2 het getal 14 en in cel A3 het getal 14,1. Selecteer de cellen A2 en A3 en sleep het plusje dat op de rand van het gemarkeerde gebied verschijnt over de A-kolom, tot de waarde 26 (in cel A102) verschijnt (we kunnen dit overigens ook doen met **Doorvoeren** onder **Bewerken**).
- Ga nu in cel B2 staan en activeer de functie NORM.VERD (dit kan natuurlijk ook onder de f_x-knop). Voer voor X cel A2 in en geef de waarden 20 respectievelijk 2 aan Gemid-

delde en Standaardafwijking. Voer bij Cumulatief nu ONWAAR in. Na OK verschijnt in cel B2 het getal 0,002215924.

- We kunnen de rest van de B-kolom invullen door het plusje op de rand van cel B2 naar beneden te slepen.
- Markeer nu het gebied A1:B102. Druk op de knop van de Grafiek-wizard en selecteer Spreiding (puntendiagram). We krijgen nu een punten-grafiek zonder verbindingsstukjes. De opmaak kan wellicht nog wat aangepast worden, maar het resultaat moet er ongeveer als volgt uitzien:

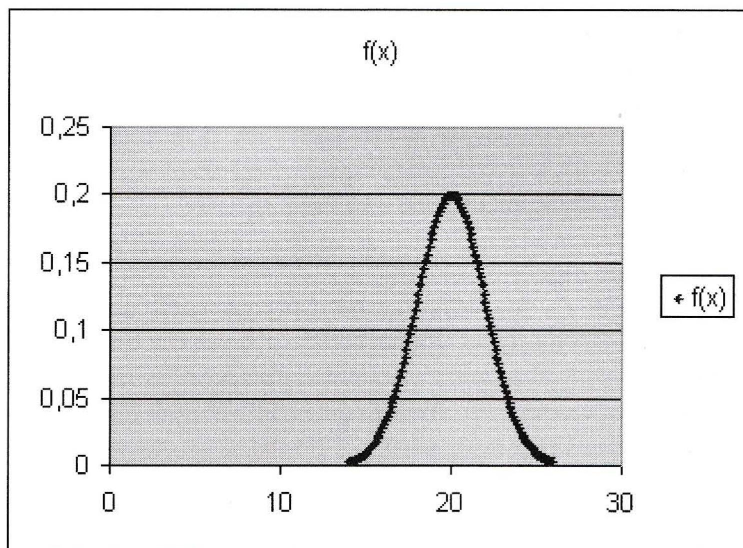


Fig. A19

A.3.6 De negatief-exponentiële verdeling

Voor de negatief-exponentiële verdeling beschikt EXCEL over de functie EXPON.VERD. Hiermee is de cumulatieve kans $P(X < x)$ te berekenen (Cumulatief = WAAR), of de kansdichtheid zelf (Cumulatief = ONWAAR).

In voorbeeld 12 uit hoofdstuk 6 wordt de kans gevraagd dat een negatief-exponentieel verdeelde kansvariabele met een parameter $\lambda = \frac{4}{3}$ meer dan 1 maar minder dan 2 bedraagt.

We berekenen deze kans door EXPON.VERD(2; 4/3; ONWAAR) en EXPON.VERD(1; 4/3; ONWAAR) van elkaar af te trekken (resultaat: 0,930516549-0,736402862=0,194113687).

Op dezelfde manier als bij de normale verdeling kan ook de kansdichtheid in beeld worden gebracht.

A.4 Schatten en toetsen

Bij het schatten worden op basis van steekproefresultaten betrouwbaarheidsintervallen voor de parameters van een populatie geconstrueerd. Wanneer een betrouwbaarheidsinterval voor het populatiegemiddelde μ geconstrueerd moet worden, kunnen we, als de standaardafwijking σ bekend is, de functie BETROUWBAARHEID gebruiken.

Voorbeeld: stel dat van een populatie $\sigma = 10$ bedraagt. We nemen een steekproef van 16 stuks en rekenen het steekproefgemiddelde uit. Dit levert bijvoorbeeld de waarde $\bar{x} = 110$ (we gebruiken de getallen uit hoofdstuk 8, het eerste voorbeeld). BETROUWBAARHEID vraagt om een waarde van α (het significantieniveau). Kies hiervoor bijvoorbeeld 0,05. Verder wordt de standaardafwijking gevraagd (=10) en de steekproefgrootte (=16). Het resultaat is 4,899902706. Dit wil zeggen dat het populatiegemiddelde met $(100 \times (1 - 0{,}05) = 95\%$ 'zekerheid' zich bevindt op het interval met:

ondergrens = $110 - \frac{1}{2} \times 4{,}899902706 \approx 105{,}1$ en

bovengrens = $110 + \frac{1}{2} \times 4{,}899902706 \approx 114{,}9$.

Voor de uitleg van de achtergrond-theorie verwijzen we naar de tekst in hoofdstuk 8.

A.4.1 De t-verdeling

Wanneer een interval voor μ geschat moet worden met onbekende standaardafwijking, dient de t-verdeling te worden toegepast. Bij het construeren van een betrouwbaarheidsinterval zal meestal de functie T.INV gebruikt worden. Deze berekent bij een bepaalde onbetrouwbaarheid α tweezijdig verdeeld ($\frac{1}{2}\alpha$ links en $\frac{1}{2}\alpha$ rechts) de benodigde *positieve* t-waarde, afhankelijk van het aantal vrijheidsgraden.

Voorbeeld: zie voorbeeld 2 uit hoofdstuk 8. De benodigde t-waarde berekenen we met T.INV(0,05; 15), met als resultaat 2,131450856. Hiermee kan het betrouwbaarheidsinterval worden geconstrueerd.

Willen we een rechteroverschrijdingskans voor een t-verdeelde variabele bepalen, dan gebruiken we de functie T.VERD. We kunnen daarin opgeven of we een tweezijdig verdeelde kans willen bepalen of een éénzijdig verdeelde kans.

A.4.2 De χ^2-verdeling

Voor het gebruik van de chi-kwadraatverdeling bij het construeren van betrouwbaarheidsintervallen zal in het algemeen de functie CHI.KWADRAAT.INV gebruikt worden. Deze rekent bij gegeven rechteroverschrijdingskans en het aantal vrijheidsgraden de bijbehorende χ^2-waarde uit.

A.4.3 Toetsen

Voor het verrichten van verschil- of vergelijkingstoetsen heeft EXCEL een aantal standaardmethodes onder **Extra**, **Gegevensanalyse**. Voor de u-toets (toetsen van gemiddelde met bekende standaardafwijking), de t-toets (toetsen van gemiddelde met onbekende standaardafwijking), de χ^2-toets (toetsen van varianties en verdelingen) en de F-toets (toetsen van

gelijkheid van varianties) kunnen we de bekende verdelingen gebruiken. We zullen als voorbeeld een t-toets uitvoeren. Het principe van het toetsen met EXCEL wordt daarmee geïllustreerd.
De gegevens nemen we uit voorbeeld 9, hoofdstuk 9. Voor de nulhypothese is geformuleerd: $\mu = 16$, voor de alternatieve hypothes geldt: $\mu > 16$. Voor de onbetrouwbaarheid geldt $\alpha = 0{,}05$. De standaardafwijking van de populatie is onbekend. Om de standaardafwijking te kunnen schatten, wordt een steekproef van 5 stuks genomen, met als uitkomsten: 15,7 - 16,3 - 16,5 -15,9 en 16,3.

We stellen voor als volgt te werk te gaan. Typ in cel A1 xi en daaronder de 5 meetgegevens. Typ in cel C1 'nulhypothese' en daaronder (C2) 'mu', (C3) 'alpha' en (C4) '1-zijdig- of 2-zijdig'? Voor mu kan 16 worden ingevuld (cel E2), voor alpha 0,05 (cel E3) en daaronder 1 (E4).
Typ in cel A7 'steekproefgegevens' en daaronder 's', 'xgem' en 'n'. s kan worden uitgerekend met STDEV (markeren A2:A6), xgem met GEMIDDELDE en n met AANTAL (steeds op cel A2:A6).
Voor de toetsingsvariabele geldt:

$$T = \frac{xgem - mu}{\frac{s}{\sqrt{n}}}$$

Typ in cel C7 'toetsingsvariabele' en daaronder 'standaardfout' (C8), 't-waarde' (C9), 'kritieke waarde' (C10) en 'overschrijdingskans' (C11). Onder standaardfout wordt verstaan $\frac{s}{\sqrt{n}}$. Deze berekenen we in cel E8 (typ in het formulevenster een =-teken, gevolgd door B8/WORTEL(B10)). De t-waarde wordt berekend in cel D9 (deze is dus (B10-E2)/E8). De kritieke waarde is de t-waarde bij een rechteroverschrijdingskans (1-zijdig) van $alpha = 0{,}05$, met $n - 1$ vrijheidsgraden, roep in cel E10 dus de functie T.INV aan en vul de betreffende gegevens in.

	A	B	C	D	E
1	xi		nulhypothese		
2	15,7		mu		16
3	16,3		alpha		0,05
4	16,5		1-zijdig of 2-zijdig?		1
5	15,9				
6	16,3				
7	steekproefgegevens		toetsingsvariabele		
8	s	0,328634	standaardfout		0,146969
9	xgem	16,14	t-waarde		0,952579
10	n	5	kritieke waarde		2,776451
11			overschrijdingskans		0,197373

Fig. A20

We kunnen nu al zeggen dat de nulhypothese niet wordt verworpen omdat de kritieke waarde (2,776451) groter is dan de t-waarde, berekend in cel E9.
Een andere manier is de bij de gevonden t-waarde behorende rechteroverschrijdingskans (één-zijdig) te berekenen. Dat doen we in cel E11 met de functie T.VERD en invulling van de gegevens. Het antwoord (0,197373) is veel groter dan alpha (=0,05), zodat de nulhypothese niet wordt verworpen.

A.5 Regressieanalyse en correlatie

Met de functie LIJNSCH (lijnschatten) geeft EXCEL de lijn die op basis van het kleinste kwadratenkriterium een ingevoerd aantal meetpunten het beste benadert. Bovendien kan met deze functie meervoudige regressie worden uitgevoerd. Verder wordt, indien gewenst, een aantal regressiegegevens opgeleverd, ook de correlatiecoëfficiënt. Ingevoerd moeten worden het bereik van de y-waarden (Y-bekend, afhankelijk), het bijbehorende bereik van de x-waarden (X-bekend, onafhankelijk). We zullen dit demonstreren voor voorbeeld 1 uit hoofdstuk 10.
De meetgegevens plaatsen we, voorzien van labels x respectievelijk y in de A en de B-kolom. Zo worden de cellen A1:A7 en B1:B7 gevuld. Selecteer voor de uitvoer een gebied met cellen waarvan het aantal kolommen gelijk is aan het aantal onafhankelijke variabelen plus één (voor het intercept b). In dit geval dus 1+1=2 kolommen. Wanneer we alleen de coëfficiënten van de lijn willen weten, kan het aantal rijen van het geselecteerde gebied beperkt blijven tot 1.
Met CORRELATIE kan de correlatiecoëfficiënt berekend worden tussen twee groepen gegevens.
Onder **Gegevensanalyse** bevindt zich een veel meer uitgebreid stuk gereedschap REGRESSIE waarmee Regressieanalyse kan worden verricht. We laten het aan de lezer over om hier zelf mee te experimenteren.

Ten slotte: Deze bijlage is bedoeld als aanzet tot het gebruik van EXCEL. Er zijn veel meer mogelijkheden dan tot dusver is gesuggereerd. De help-functie bij EXCEL zal desgewenst de lezer helpen bij het beantwoorden van zijn of haar vragen.

Bijlage B Tabellen

B1 Rechteroverschrijdingskansen in de (standaardnormale)

U-verdeling: $P(U > u) = \frac{1}{\sqrt{2\pi}} \int_u^{\infty} e^{-\frac{1}{2}t^2} dt$

u	0	1	2	3	4	5	6	7	8	9
0,0	5000	4960	4920	4880	4840	4801	4761	4721	4681	4641
0,1	4602	4562	4522	4483	4443	4404	4364	4325	4286	4247
0,2	4207	4168	4129	4090	4052	4013	3974	3936	3897	3859
0,3	3821	3783	3745	3707	3669	3632	3594	3557	3520	3483
0,4	3446	3409	3372	3336	3300	3264	3228	3192	3156	3121
0,5	3085	3050	3015	2981	2946	2912	2877	2843	2810	2776
0,6	2743	2709	2676	2643	2611	2578	2546	2514	2483	2451
0,7	2420	2389	2358	2327	2296	2266	2236	2206	2177	2148
0,8	2119	2090	2061	2033	2005	1977	1949	1922	1894	1867
0,9	1841	1814	1788	1762	1736	1711	1685	1660	1635	1611
1,0	1587	1562	1539	1515	1492	1469	1446	1423	1401	1379
1,1	1357	1335	1314	1292	1271	1251	1230	1210	1190	1170
1,2	1151	1131	1112	1093	1075	1056	1038	1020	1003	0985
1,3	0968	0951	0934	0918	0901	0885	0869	0853	0838	0823
1,4	0808	0793	0778	0764	0749	0735	0721	0708	0694	0681
1,5	0668	0655	0643	0630	0618	0606	0594	0582	0571	0559
1,6	0548	0537	0526	0516	0505	0495	0485	0475	0465	0455
1,7	0446	0436	0427	0418	0409	0401	0392	0384	0375	0367
1,8	0359	0351	0344	0336	0329	0322	0314	0307	0301	0294
1,9	0287	0281	0274	0268	0262	0256	0250	0244	0239	0233
2,0	0228	0222	0217	0210	0207	0202	0197	0192	0188	0183
2,1	0179	0174	0170	0166	0162	0158	0154	0150	0146	0143
2,2	0139	0136	0132	0129	0125	0122	0119	0116	0113	0110
2,3	0107	0104	0102	0099	0096	0094	0091	0089	0087	0084
2,4	0082	0080	0078	0075	0073	0071	0069	0068	0066	0064
2,5	0062	0060	0059	0057	0055	0054	0052	0051	0049	0048
2,6	0047	0045	0044	0043	0041	0040	0039	0038	0037	0036
2,7	0035	0034	0033	0032	0031	0030	0029	0028	0027	0026
2,8	0026	0025	0024	0023	0023	0022	0021	0021	0020	0019
2,9	0019	0018	0018	0017	0016	0016	0015	0015	0014	0014
3,0	0013	0013	0013	0012	0012	0011	0011	0011	0010	0010
3,1	0010	0009	0009	0009	0008	0008	0008	0008	0007	0007
3,2	0007	0007	0006	0006	0006	0006	0006	0005	0005	0005
3,3	0005	0005	0005	0004	0004	0004	0004	0004	0004	0003
3,4	0003	0003	0003	0003	0003	0003	0003	0003	0003	0002

N.B. De rechteroverschrijdingskansen zijn met 10^4 vermenigvuldigd.

B2 Binomiale verdelingen voor enkele waarden van n en p: $P(K = k) = \binom{n}{k} p^k.(1-p)^{n-k}$

		p											
n	k	0,01	0,05	0,10	0,15	0,20	0,25	0,30	1/3	0,35	0,40	0,45	0,50
2	0	9801	9025	8100	7225	6400	5625	4900	4444	4225	3600	3025	2500
	1	0198	0950	1800	2550	3200	3750	4200	4444	4550	4800	4950	5000
	2	0001	0025	0100	0225	0400	0625	0900	1111	1225	1600	2025	2500
3	0	9703	8574	7290	6141	5120	4219	3430	2963	2746	2160	1664	1250
	1	0294	1354	2430	3251	3840	4219	4410	4444	4436	4320	4084	3750
	2	0003	0071	0270	0574	0960	1406	1890	2222	2389	2880	3341	3750
	3	0000	0001	0010	0034	0080	0156	0270	0370	0429	0640	0911	1250
4	0	9606	8145	6561	5220	4096	3164	2401	1975	1785	1296	0915	0625
	1	0388	1715	2916	3685	4096	4219	4116	3951	3845	3456	2995	2500
	2	0006	0135	0486	0975	1536	2109	2646	2963	3105	3456	3675	3750
	3	0000	0005	0036	0115	0256	0469	0756	0988	1115	1536	2005	2500
	4	0000	0000	0001	0005	0016	0039	0081	0123	0150	0256	0410	0625
5	0	9510	7738	5905	4437	3277	2373	1681	1317	1160	0778	0503	0312
	1	0480	2036	3280	3915	4096	3955	3602	3292	3124	2592	2059	1562
	2	0010	0214	0729	1382	2048	2637	3087	3292	3364	3456	3369	3125
	3	0000	0011	0081	0244	0512	0879	1323	1646	1811	2304	2757	3125
	4	0000	0000	0004	0022	0064	0146	0284	0412	0488	0768	1128	1562
	5	0000	0000	0000	0001	0003	0010	0024	0041	0053	0102	0185	0312
6	0	9415	7351	5314	3771	2621	1780	1176	0878	0754	0467	0277	0156
	1	0571	2321	3543	3993	3932	3560	3025	2634	2437	1866	1359	0938
	2	0014	0305	0984	1762	2458	2966	3241	3292	3280	3110	2780	2344
	3	0000	0021	0146	0415	0819	1318	1852	2195	2355	2765	3032	3125
	4	0000	0001	0012	0055	0154	0330	0595	0823	0951	1382	1861	2344
	5	0000	0000	0001	0004	0015	0044	0102	0165	0205	0369	0609	0938
	6	0000	0000	0000	0000	0001	0002	0007	0014	0018	0041	0083	0156
7	0	9321	6983	4783	3206	2097	1335	0824	0585	0490	0280	0152	0078
	1	0659	2573	3720	3960	3670	3115	2471	2048	1848	1306	0872	0547
	2	0020	0406	1240	2097	2753	3115	3177	3073	2985	2613	2140	1641
	3	0000	0036	0230	0617	1147	1730	2269	2561	2679	2903	2918	2734
	4	0000	0002	0026	0109	0287	0577	0972	1280	1442	1935	2388	2734
	5	0000	0000	0002	0012	0043	0115	0250	0384	0466	0774	1172	1641
	6	0000	0000	0000	0001	0004	0013	0036	0064	0084	0172	0320	0547
	7	0000	0000	0000	0000	0000	0001	0002	0005	0006	0016	0037	0078

B2 (vervolg)

n	k	p											
		0,01	0,05	0,10	0,15	0,20	0,25	0,30	1/3	0,35	0,40	0,45	0,50
8	0	9227	6634	4305	2725	1678	1001	0576	0390	0319	0168	0084	0039
	1	0746	2793	3826	3847	3355	2670	1977	1561	1373	0896	0548	0312
	2	0026	0515	1488	2376	2936	3115	2965	2731	2587	2090	1569	1094
	3	0001	0054	0331	0839	1468	2076	2541	2731	2786	2787	2568	2188
	4	0000	0004	0046	0185	0459	0865	1361	1707	1875	2322	2627	2734
	5	0000	0000	0004	0026	0092	0231	0467	0683	0808	1239	1719	2188
	6	0000	0000	0000	0002	0011	0038	0100	0171	0217	0413	0703	1094
	7	0000	0000	0000	0000	0001	0004	0012	0024	0033	0079	0164	0312
	8	0000	0000	0000	0000	0000	0000	0001	0002	0002	0007	0017	0039
9	0	9135	6302	3874	2316	1342	0751	0404	0260	0207	0101	0046	0020
	1	0830	2985	3874	3679	3020	2253	1556	1171	1004	0605	0339	0176
	2	0034	0629	1722	2597	3020	3003	2668	2341	2162	1612	1110	0703
	3	0001	0077	0446	1069	1762	2336	2668	2731	2716	2508	2119	1641
	4	0000	0006	0074	0283	0661	1168	1715	2048	2194	2508	2600	2461
	5	0000	0000	0008	0050	0165	0389	0735	1024	1181	1672	2128	2461
	6	0000	0000	0001	0006	0028	0087	0210	0341	0424	0743	1160	1641
	7	0000	0000	0000	0000	0003	0012	0039	0073	0098	0210	0407	0703
	8	0000	0000	0000	0000	0000	0001	0004	0009	0013	0035	0083	0176
	9	0000	0000	0000	0000	0000	0000	0000	0001	0001	0003	0008	0020
10	0	9044	5987	3487	1969	1074	0563	0282	0173	0135	0060	0025	0010
	1	0914	3151	3874	3474	2684	1877	1211	0867	0725	0403	0207	0098
	2	0042	0746	1937	2759	3020	2816	2335	1951	1757	1209	0763	0439
	3	0001	0105	0574	1298	2013	2503	2668	2601	2522	2150	1665	1172
	4	0000	0010	0112	0401	0881	1460	2001	2276	2377	2508	2384	2051
	5	0000	0001	0015	0085	0264	0584	1029	1366	1536	2007	2340	2431
	6	0000	0000	0001	0012	0055	0162	0368	0569	0689	1115	1596	2051
	7	0000	0000	0000	0001	0008	0031	0090	0163	0212	0425	0746	1172
	8	0000	0000	0000	0000	0001	0004	0014	0030	0043	0106	0229	0439
	9	0000	0000	0000	0000	0000	0000	0001	0003	0005	0016	0042	0098
	10	0000	0000	0000	0000	0000	0000	0000	0000	0000	0001	0003	0010

N.B. De kansen zijn met 10^4 vermenigvuldigd.

B3 De enkelvoudige Poisson-verdeling: $P(K = k) = \frac{m^k e^{-m}}{k!}$

m	k							
	0	1	2	3	4	5	6	7
0,05	0,951	0,048	0,001					
0,10	0,905	0,090	0,005					
0,15	0,861	0,129	0,009	0,001				
0,20	0,819	0,163	0,017	0,001				
0,25	0,779	0,195	0,024	0,002				
0,30	0,741	0,222	0,033	0,004				
0,35	0,705	0,246	0,043	0,006				
0,40	0,670	0,268	0,054	0,007	0,001			
0,45	0,638	0,287	0,064	0,010	0,001			
0,50	0,607	0,303	0,076	0,012	0,002			
0,55	0,577	0,317	0,088	0016	0,002			
0,60	0,549	0,329	0,099	0’020	0,003			
0,65	0,522	0,339	0,111	0,024	0,003	0,001		
0,70	0,497	0,347	0,122	0,028	0,005	0,001		
0,75	0,472	0,355	0,132	0,034	0,006	0,001		
0,80	0,449	0,360	0,144	0,038	0,008	0,001		
0,85	0,427	0,364	0,154	0,044	0,009	0,002		
0,90	0,407	0,365	0,165	0,050	0,011	0,002		
0,95	0,387	0,367	0,175	0,055	0,013	0,003		
1,00	0,368	0,368	0,184	0,061	0,015	0,003	0,001	
1,1	0,333	0,366	0,201	0,074	0,021	0,004	0,001	
1,2	0,301	0,362	0,216	0,087	0,026	0,006	0,002	
1,3	0,273	0,354	0,230	0,100	0,032	0,009	0,002	
1,4	0,247	0,345	0,241	0,113	0,040	0,011	0,002	0,001
1,5	0,223	0,335	0,251	0,125	0,047	0,015	0,003	0,001

B3 (vervolg)

m	k											
	0	1	2	3	4	5	6	7	8	9	10	11
1,6	0,202	0,323	0,258	0,138	0,055	0,018	0,005	0,001				
1,7	0,183	0,310	0,264	0,150	0,063	0,022	0,006	0,002				
1,8	0,165	0,298	0,268	0,160	0,073	0,026	0,007	0,002	0,001			
1,9	0,150	0,284	0,270	0,171	0,081	0,031	0,010	0,002	0,001			
2,0	0,135	0,271	0,271	0,180	0,090	0,036	0,012	0,004	0,001			
2,2	0,111	0,244	0,268	0,196	0,109	0,047	0,018	0,005	0,002			
2,4	0,091	0,217	0,262	0,209	0,125	0,060	0,024	0,009	0,002	0,001		
2,6	0,074	0,193	0,251	0,218	0,141	0,074	0,032	0,012	0,004	0,001		
2,8	0,061	0,170	0,238	0,223	0,156	0,087	0,041	0,016	0,006	0,001	0,001	
3,0	0,050	0,149	0,224	0,224	0,168	0,101	0,050	0,022	0,008	0,003	0,001	
3,2	0,041	0,130	0,209	0,223	0,178	0,114	0,060	0,028	0,011	0,004	0,002	
3,4	0,033	0,114	0,193	0,218	0,186	0,127	0,071	0,035	0,015	0,005	0,002	0,001
3,6	0,027	0,099	0,177	0,212	0,191	0,138	0,083	0,042	0,019	0,008	0,003	0,001
3,8	0,022	0,085	0,162	0,204	0,195	0,148	0,093	0,051	0,024	0,010	0,004	0,001
4,0	0,018	0,074	0,146	0,195	0,196	0,156	0,104	0,060	0,030	0,013	0,005	0,002
4,5	0,011	0,050	0,113	0,168	0,190	0,171	0,128	0,082	0,047	0,023	0,010	0,004
5,0	0,007	0,033	0,085	0,140	0,175	0,176	0,146	0,105	0,065	0,036	0,018	0,009
5,5	0,005	0,022	0,061	0,113	0,157	0,171	0,157	0,124	0,084	0,052	0,029	0,014
6,0	0,002	0,015	0,045	0,089	0,134	0,161	0,160	0,138	0,103	0,069	0,041	0,023
6,5	0,001	0,010	0,032	0,069	0,112	0,146	0,157	0,146	0,119	0,085	0,056	0,033
7,0	0,001	0,006	0,023	0,052	0,091	0,128	0,149	0,149	0,130	0,101	0,071	0,046
7,5	0,001	0,004	0,016	0,038	0,074	0,109	0,127	0,146	0,137	0,115	0,086	0,058
8,0	0,000	0,003	0,011	0,028	0,058	0,091	0,122	0,140	0,140	0,124	0,099	0,072
8,5	0,000	0,002	0,007	0,021	0,044	0,076	0,106	0,130	0,137	0,130	0,110	0,086
9,0	0,000	0,001	0,005	0,015	0,034	0,061	0,091	0,117	0,132	0,131	0,119	0,097
9,5	0,000	0,001	0,003	0,011	0,025	0,049	0,076	0,104	0,123	0,130	0,123	0,107
10,0	0,000	0,000	0,003	0,007	0,019	0,038	0,063	0,090	0,113	0,125	0,125	0,114

m	k										
	12	13	14	15	16	17	18	19	20	21	22
1,6											
1,7											
1,8											
1,9											
2,0											
2,2											
2,4											
2,6											
2,8											
3,0											
3,2											
3,4											
3,6											
3,8	0,001										
4,0	0,001										
4,5	0,002	0,001									
5,0	0,003	0,001	0,001								
5,5	0,006	0,003	0,001	0,001							
6,0	0,011	0,005	0,003	0,001	0,001						
6,5	0,018	0,009	0,004	0,002	0,001						
7,0	0,027	0,014	0,007	0,004	0,001	0,001					
7,5	0,037	0,020	0,012	0,006	0,002	0,001	0,001				
8,0	0,048	0,030	0,017	0,009	0,004	0,002	0,001				
8,5	0,060	0,040	0,024	0,013	0,007	0,004	0,002	0,001			
9,0	0,073	0,050	0,033	0,019	0,011	0,006	0,003	0,001	0,001		
9,5	0,084	0,062	0,042	0,027	0,015	0,009	0,005	0,002	0,001	0,001	
10,0	0,095	0,072	0,053	0,034	0,022	0,013	0,007	0,004	0,001	0,001	0,001

B4 De cumulatieve Poisson-verdeling: $P(K \leq c) = \sum_{k=0}^{c} \frac{m^k e^{-m}}{k!}$

$= P(K = 0) + P(K = 1) + P(K = 2) ... + P(K = c)$

m	c											
	0	1	2	3	4	5	6	7	8	9	10	11
0,05	0,951	0,999	1,000									
0,10	0,905	0,995	1,000									
0,15	0,861	0,990	0,999	1,000								
0,20	0,819	0,982	0,999	1,000								
0,25	0,779	0,974	0,998	1,000								
0,30	0,741	0,963	0,996	1,000								
0,35	0,705	0,951	0,994	1,000								
0,40	0,670	0,938	0,992	0,999	1,000							
0,45	0,638	0,925	0,989	0,999	1,000							
0,50	0,607	0,910	0,986	0,998	1,000							
0,55	0,577	0,894	0,982	0,998	1,000							
0,60	0,549	0,878	0,977	0,997	1,000							
0,65	0,522	0,861	0,972	0,996	0,999	1,000						
0,70	0,497	0,844	0,966	0,994	0,999	1,000						
0,75	0,472	0,827	0,959	0,993	0,999	1,000						
0,80	0,449	0,809	0,953	0,991	0,999	1,000						
0,85	0,427	0,791	0,945	0,989	0,998	1,000						
0,90	0,407	0,772	0,937	0,987	0,998	1,000						
0,95	0,387	0,754	0,929	0,984	0,997	1,000						
1,00	0,368	0,736	0,920	0,981	0,996	0,999	1,000					
1,1	0,333	0,699	0,900	0,974	0,995	0,999	1,000					
1,2	0,301	0,663	0,879	0,966	0,992	0,998	1,000					
1,3	0,273	0,627	0,857	0,957	0,989	0,998	1,000					
1,4	0,247	0,592	0,833	0,946	0,986	0,997	0,999	1,000				
1,5	0,223	0,558	0,809	0,934	0,981	0,996	0,999	1,000				

B4 (vervolg)

m	c											
	0	1	2	3	4	5	6	7	8	9	10	11
1,6	0,202	0,525	0,783	0,921	0,976	0,994	0,999	1,000				
1,7	0,183	0,493	0,757	0,907	0,970	0,992	0,998	1,000				
1,8	0,165	0,463	0,731	0,891	0,964	0,990	0,997	0,999	1000			
1,9	0,150	0,434	0,704	0,875	0,956	0,987	0,997	0,999	1,000			
2,0	0,135	0,406	0,677	0,857	0,947	0,983	0,995	0,999	1,000			
2,2	0,111	0,355	0,623	0,819	0,928	0,975	0,993	0,998	1,000			
2,4	0,091	0,308	0,570	0,779	0,904	0,964	0,988	0,997	0,999	1,000		
2,6	0,074	0,267	0,518	0,736	0,877	0,951	0,983	0,995	0,999	1,000		
2,8	0,061	0,231	0,469	0,692	0,848	0,935	0,976	0,992	0,998	0,999	1,000	
3,0	0,050	0,199	0,423	0,647	0,815	0,916	0,966	0,988	0,996	0,999	1,000	
3,2	0,041	0,171	0,380	0,603	0,781	0,895	0,955	0,983	0,994	0,998	1,000	
3,4	0,033	0147	0,340	0,558	0,744	0,871	0,942	0,977	0,992	0,997	0,999	1,000
3,6	0,027	0,126	0,303	0,515	0,706	0,844	0,927	0,969	0,988	0,996	0,999	1,000
3,8	0,022	0,107	0,269	0,473	0,668	0,816	0,909	0,960	0,984	0,994	0,998	0,999
4,0	0,018	0,092	0,238	0,433	0,629	0,785	0,889	0,949	0,979	0,992	0,997	0,999
4,5	0,011	0,061	0,174	0,342	0,532	0,703	0,830	0,913	0,960	0,983	0,993	0,997
5,0	0,007	0,040	0,125	0,265	0,440	0,616	0,762	0,867	0,932	0,968	0,986	0,995
5,5	0,005	0,027	0,088	0,201	0,358	0,529	0,686	0,810	0,894	0,946	0,975	0,989
6,0	0,002	0,017	0,062	0,151	0,285	0,446	0,606	0,744	0,847	0,916	0,957	0,980
6,5	0,001	0,011	0,043	0,112	0,224	0,370	0,527	0,673	0,792	0,877	0,933	0,966
7,0	0,001	0,007	0,030	0,082	0,173	0,301	0,450	0,599	0,729	0,830	0,901	0,947
7,5	0,001	0,005	0,021	0,059	0,133	0,242	0,379	0,525	0,662	0,777	0,863	0,921
8,0	0,000	0,003	0,014	0,042	0,100	0,191	0,313	0,453	0,593	0,717	0,816	0,888
8,5	0,000	0,002	0,009	0,030	0,074	0,150	0,256	0,386	0,523	0,653	0,703	0,849
9,0	0,000	0,001	0,006	0,021	0,055	0,116	0,207	0,324	0,456	0,587	0,706	0,803
9,5	0,000	0,001	0,004	0,015	0,040	0,089	0,165	0,269	0,392	0,522	0,645	0,752
10,0	0,000	0,000	0,003	0,010	0,029	0,067	0,130	0,220	0,333	0,458	0,583	0,697

m	c										
	12	13	14	15	16	17	18	19	20	21	22
1,6											
1,7											
1,8											
1,9											
2,0											
2,2											
2,4											
2,6											
2,8											
3,0											
3,2											
3,4											
3,6											
3,8	1,000										
4,0	1,000										
4,5	0,999	1,000									
5,0	0,998	0,999	1,000								
5,5	0,995	0,998	0,999	1,000							
6,0	0,991	0,996	0,999	0,999	1,000						
6,5	0,984	0,993	0,997	0,999	1,000						
7,0	0,973	0,987	0,994	0,998	0,999	1,000					
7,5	0,958	0,978	0,990	0,996	0,998	0,999	1,000				
8,0	0,936	0,966	0,983	0,992	0,996	0,998	0,999	1,000			
8,5	0,909	0,949	0,973	0,986	0,993	0,997	0,999	0,999	1,000		
9,0	0,876	0,926	0,959	0,978	0,989	0,995	0,998	0,999	1,000		
9,5	0,836	0,898	0,940	0,967	0,982	0,991	0,996	0,998	0,999	1,000	
10,0	0,792	0,864	0,917	0,951	0,973	0,986	0,993	0,997	0,998	0,999	1,000

B5 Rechter kritieke waarden in de T-verdeling: waarden $t_\nu(\alpha)$ van T

ν	α					
	0,10	0,05	0,025	0,01	0,005	0,0005
1	3,078	6,314	12,706	31,821	63,657	636,619
2	1,886	2,920	4,303	6,965	9,925	31,598
3	1,638	2,353	3,182	4,541	5,841	12,924
4	1,533	2,132	2,776	3,747	4,604	8,610
5	1,476	2,015	2,571	3,365	4,032	6,869
6	1,440	1,943	2,447	3,143	3,707	5,959
7	1,415	1,895	2,365	2,998	3,499	5,408
8	1,397	1,860	2,306	2,896	3,355	5,041
9	1,383	1,833	2,262	2,821	3,250	4,781
10	1,372	1,812	2,228	2,764	3,169	4,587
11	1,363	1,796	2,201	2,718	3,106	4,437
12	1,356	1,782	2,179	2,681	3,055	4,318
13	1,350	1,771	2,160	2,650	3,012	4,221
14	1,345	1,761	2,145	2,624	2,977	4,140
15	1,341	1,753	2,131	2,602	2,947	4,073
16	1,337	1,746	2,120	2,583	2,921	4,015
17	1,333	1,740	2,110	2,567	2,898	3,965
18	1,330	1,734	2,101	2,552	2,878	3,922
19	1,328	1,729	2,093	2,539	2,861	3,883
20	1,325	1,725	2,086	2,528	2,845	3,850
21	1,323	1,721	2,080	2,518	2,831	3,819
22	1,321	1,717	2,074	2,508	2,819	3,792
23	1,319	1,714	2,069	2,500	2,807	3,768
24	1,318	1,711	2,064	2,492	2,797	3,745
25	1,316	1,708	2,060	2,485	2,787	3,725
26	1,315	1,705	2,056	2,479	2,779	3,707
27	1,314	1,703	2,052	2,473	2,771	3,690
28	1,313	1,701	2,048	2,467	2,763	3,674
29	1,311	1,699	2,045	2,462	2,756	3,659
30	1,310	1,697	2,042	2,457	2,750	3,646
40	1,303	1,684	2,021	2,423	2,704	3,551
60	1,296	1,671	2,000	2,390	2,660	3,460
120	1,289	1,658	1,980	2,358	2,617	3,373
∞	1,282	1,645	1,960	2,326	2,576	3,291

B6 Rechter kritieke waarden in de χ^2-verdeling: waarden van $\chi^2_\nu(\alpha)$

ν	α								
	0,99	0,975	0,95	0,90	0,50	0,10	0,05	0,025	0,01
1	0,000	0,001	0,004	0,015	0,455	2,71	3,84	5,02	6,64
2	0,020	0,051	0,103	0,211	1,386	4,61	5,99	7,38	9,21
3	0,115	0,216	0,352	0,584	2,366	6,25	7,82	9,35	11,34
4	0,297	0,484	0,711	1,064	3,357	7,78	9,49	11,14	13,28
5	0,554	0,831	1,145	1,610	4,351	9,24	11,07	12,83	15,09
6	0,872	1,237	1,635	2,204	5,35	10,65	12,59	14,45	16,81
7	1,239	1,690	2,167	2,833	6,35	12,02	14,07	16,01	18,48
8	1,646	2,180	2,733	3,490	7,34	13,36	15,51	17,53	20,09
9	2,088	2,700	3,325	4,168	8,34	14,68	16,92	19,02	21,67
10	2,558	3,247	3,940	4,865	9,34	15,99	18,31	20,48	23,21
11	3,05	3,82	4,57	5,58	10,34	17,28	19,68	21,92	24,73
12	3,57	4,40	5,23	6,30	11,34	18,55	21,03	23,34	26,22
13	4,11	5,01	5,89	7,04	12,34	19,81	22,36	24,74	27,69
14	4,66	5,63	6,57	7,79	13,34	21,06	23,69	26,12	29,14
15	5,23	6,26	7,26	8,55	14,34	22,31	25,00	27,49	30,58
16	5,81	6,91	7,96	9,31	15,34	23,54	26,30	28,85	32,00
17	6,41	7,56	8,67	10,09	16,34	24,77	27,59	30,19	33,41
18	7,01	8,23	9,39	10,87	17,34	25,99	28,87	31,53	34,81
19	7,63	8,91	10,12	11,65	18,34	27,20	30,14	32,85	36,19
20	8,26	9,59	10,85	12,44	19,34	28,41	31,41	34,17	37,57
21	8,90	10,28	11,59	13,34	20,34	29,61	32,67	35,48	38,93
22	9,54	10,98	12,34	14,04	21,34	30,81	33,92	36,78	40,29
23	10,20	11,69	13,09	14,85	22,34	32,01	35,17	38,08	41,64
24	10,86	12,40	13,85	15,66	23,34	33,20	36,42	39,36	42,98
25	11,52	13,12	14,61	16,47	24,34	34,38	37,65	40,65	44,31
26	12,20	13,84	15,38	17,29	25,34	35,56	38,89	41,92	45,64
27	12,88	14,57	16,15	18,11	26,34	36,74	40,11	43,19	46,96
28	13,56	15,31	16,93	18,94	27,34	37,92	41,34	44,46	48,28
29	14,26	16,05	17,71	19,77	28,34	39,09	42,56	45,72	49,59
30	14,95	16,79	18,49	20,60	29,34	40,26	43,77	46,98	50,89

B7 Rechter kritieke waarden in de F-verdeling: waarden van $F_{\nu_1,\nu_2}(0{,}05)$

ν_2	ν_1									
	1	2	3	4	5	6	7	8	9	10
1	161,4	199,5	215,7	224,6	230,2	234,0	236,8	238,9	240,5	241,9
2	18,5	19,0	19,2	19,2	19,3	19,3	19,4	19,4	19,4	19,4
3	10,1	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99
120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,96	1,91
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83

B7 (vervolg)

ν_2	ν_1								
	12	15	20	24	30	40	60	120	∞
1	243,9	245,9	248,0	249,1	250,1	251,1	252,2	253,3	254,3
2	19,4	19,4	19,4	19,5	19,5	19,5	19,5	19,5	19,5
3	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,36
6	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
8	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
9	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
11	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
12	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
13	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
15	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
20	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
21	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
22	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
23	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
24	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
25	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
30	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
40	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
60	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25
∞	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00

B8 Rechter kritieke waarden in de F-verdeling: waarden van $F_{\nu_1,\nu_2}(0{,}025)$

ν_2	ν_1									
	1	2	3	4	5	6	7	8	9	10
1	647,5	799,5	864,2	899,6	921,8	937,1	948,2	956,7	963,3	968,6
2	38,5	39,0	39,2	39,2	39,3	39,3	39,4	39,4	39,4	39,4
3	17,4	16,0	15,4	15,1	14,9	14,7	14,6	14,5	14,5	14,4
4	12,2	10,6	9,98	9,60	9,36	9,20	9,07	8,98	8,90	8,84
5	10,0	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68	6,62
6	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,52	5,46
7	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,82	4,76
8	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36	4,30
9	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	4,03	3,96
10	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,78	3,72
11	6,72	5,26	4,63	4,28	4,04	3,88	3,76	3,66	3,59	3,53
12	6,55	5,10	4,47	4,12	3,89	3,73	3,61	3,51	3,44	3,37
13	6,41	4,97	4,35	4,00	3,77	3,60	3,48	3,39	3,31	3,25
14	6,30	4,86	4,24	3,89	3,66	3,50	3,38	3,29	3,21	3,15
15	6,20	4,77	4,15	3,80	3,58	3,41	3,29	3,20	3,12	3,06
16	6,12	4,69	4,08	3,73	3,50	3,34	3,22	3,12	3,05	2,99
17	6,04	4,62	4,01	3,66	3,44	3,28	3,16	3,06	2,98	2,92
18	5,98	4,56	3,95	3,61	3,38	3,22	3,10	3,01	2,93	2,87
19	5,92	4,51	3,90	3,56	3,33	3,17	3,05	2,96	2,88	2,82
20	5,87	4,46	3,86	3,51	3,29	3,13	3,01	2,91	2,84	2,77
21	5,83	4,42	3,82	3,48	3,25	3,09	2,97	2,87	2,80	2,73
22	5,79	4,38	3,78	3,44	3,22	3,05	2,93	2,84	2,76	2,70
23	5,75	4,35	3,75	3,41	3,18	3,02	2,90	2,81	2,73	2,67
24	5,72	4,32	3,72	3,38	3,15	2,99	2,87	2,78	2,70	2,64
25	5,69	4,29	3,69	3,35	3,13	2,97	2,85	2,75	2,68	2,61
30	5,57	4,18	3,59	3,25	3,03	2,87	2,75	2,65	2,57	2,51
40	5,42	4,05	3,46	3,13	2,90	2,74	2,62	2,53	2,45	2,39
60	5,29	3,93	3,34	3,01	2,79	2,63	2,51	2,41	2,33	2,27
120	5,15	3,80	3,23	2,89	2,67	2,52	2,39	2,30	2,22	2,16
∞	5,02	3,69	3,12	2,79	2,57	2,41	2,29	2,19	2,11	2,05

B8 (vervolg)

ν_2	ν_1								
	12	15	20	24	30	40	60	120	∞
1	976,7	984,9	993,1	997,2	1001	1006	1010	1014	1018
2	39,4	39,4	39,4	39,5	39,5	39,5	39,5	39,5	39,5
3	14,3	14,3	14,2	14,1	14,1	14,0	14,0	13,9	13,9
4	8,75	8,66	8,56	8,51	8,46	8,41	8,36	8,31	8,26
5	6,52	6,43	6,33	6,28	6,23	6,18	6,12	6,07	6,02
6	5,37	5,27	5,17	5,12	5,07	5,01	4,96	4,90	4,85
7	4,67	4,57	4,47	4,42	4,36	431	4,25	4,20	4,14
8	4,20	4,10	4,00	3,95	3,89	3,84	3,78	3,73	3,67
9	3,87	3,77	3,67	3,61	3,56	3,51	3,45	3,39	3,33
10	3,62	3,52	3,42	3,37	3,31	3,26	3,20	3,14	3,08
11	3,43	3,33	3,23	3,17	3,12	3,06	3,00	2,94	2,88
12	3,28	3,18	3,07	3,02	2,96	2,91	2,85	2,79	2,72
13	3,15	3,05	2,95	2,89	2,84	2,78	2,72	2,66	2,60
14	3,05	2,95	2,84	2,79	2,73	2,67	2,61	2,55	2,49
15	2,96	2,86	2,76	2,70	2,64	2,59	2,52	2,46	2,40
16	2,89	2,79	2,68	2,63	2,S7	2,51	2,45	2,38	2,32
17	2,82	2,72	2,62	2,56	2,50	2,44	2,38	2,32	2,25
18	2,77	2,67	2,56	2,50	2,44	2,38	2,32	2,26	2,19
19	2,72	2,62	2,51	2,45	2,39	2,33	2,27	2,20	2,13
20	2,68	2,57	2,46	2,41	2,35	2,29	2,22	2,16	2,09
21	2,64	2,53	2,42	2,37	2,31	2,25	2,18	2,11	2,04
22	2,60	2,50	2,39	2,33	2,27	2,21	2,14	2,08	2,00
23	2,57	2,47	2,36	2,30	2,24	2,18	2,11	2,04	1,97
24	2,54	2,44	2,33	2,27	2,21	2,15	2,08	2,01	1,94
25	2,51	2,41	2,30	2,24	2,18	2,12	2,05	1,98	1,91
30	2,41	2,31	2,20	2,14	2,07	2,01	1,94	1,87	1,79
40	2,29	2,18	2,07	2,01	1,94	1,88	1,80	1,72	1,64
60	2,17	2,06	1,94	1,88	1,82	1,74	1,67	1,58	1,48
120	2,05	1,94	1,82	1,76	1,69	1,61	1,53	1,43	1,31
∞	1,94	1,83	1,71	1,64	1,57	1,48	1,39	1,27	1,00

B9 Rechter kritieke waarden in de F-verdeling: waarden van $F_{\nu_1,\nu_2}(0{,}01)$

ν_2	ν_1									
	1	2	3	4	5	6	7	8	9	10
1	4052	5000	5403	5625	5764	5859	5928	5982	6022	6056
2	98,5	99,0	99,2	99,2	99,3	99,3	99,4	99,4	99,4	99,4
3	34,1	30,8	29,5	28,7	28,2	27,9	27,7	27,5	27,3	27,2
4	21,2	18,0	16,7	16,0	15,5	15,2	15,0	14,8	14,7	14,5
5	16,3	13,3	12,1	11,4	11,0	10,7	10,5	10,3	10,2	10,1
6	13,7	10,9	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87
7	12,2	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62
8	11,3	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81
9	10,6	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26
10	10,0	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94
15	8,68	6,36	5,42	4,89	4,56	4,32	4,41	4,00	3,89	3,80
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69
17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17
25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47
∞	6,63	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32

B9 (vervolg)

ν_2	ν_1								
	12	15	20	24	30	40	60	120	∞
1	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	99,4	99,4	99,4	99,5	99,5	99,5	99,5	99,5	99,5
3	27,1	26,9	26,7	26,6	26,5	26,4	26,3	26,2	26,1
4	14,4	14,2	14,0	13,9	13,8	13,7	13,7	13,6	13,5
5	9,89	9,72	9,55	9,47	9,38	9,29	9,20	9,11	9,02
6	7,72	7,56	7,40	7,31	7,23	7,14	7,06	6,97	6,88
7	6,47	6,31	6,16	6,07	5,99	5,91	5,82	5,74	5,65
8	5,67	5,52	5,36	5,28	5,20	5,12	5,03	4,95	4,86
9	5,11	4,96	4,81	4,73	4,65	4,57	4,48	4,40	4,31
10	4,71	4,56	4,41	4,33	4,25	4,17	4,08	4,00	3,91
11	4,40	4,25	4,10	4,02	3,94	3,86	3,78	3,69	3,60
12	4,16	4,01	3,86	3,78	3,70	3,62	3,54	3,45	3,36
13	3,96	3,82	3,66	3,59	3,51	3,43	3,34	3,25	3,17
14	3,80	3,66	3,51	3,43	3,35	3,27	3,18	3,09	3,00
15	3,67	3,52	3,37	3,29	3,21	3,13	3,05	2,96	2,87
16	3,55	3,41	3,26	3,18	3,10	3,02	2,93	2,84	2,75
17	3,46	3,31	3,16	3,08	3,00	2,92	2,83	2,75	2,65
18	3,37	3,23	3,08	3,00	2,92	2,84	2,75	2,66	2,57
19	3,30	3,15	3,00	2,92	2,84	2,76	2,67	2,58	2,49
20	3,23	3,09	2,94	2,86	2,78	2,69	2,61	2,52	2,42
21	3,17	3,03	2,88	2,80	2,72	2,64	2,55	2,46	2,36
22	3,12	2,98	2,83	2,75	2,67	2,58	2,50	2,40	2,31
23	3,07	2,93	2,78	2,70	2,62	2,54	2,45	2,35	2,26
24	3,03	2,89	2,74	2,66	2,58	2,49	2,40	2,31	2,21
25	2,99	2,85	2,70	2,62	2,54	2,45	2,36	2,27	2,17
30	2,84	2,70	2,55	2,47	2,39	2,30	2,21	2,11	2,01
40	2,66	2,52	2,37	2,29	2,20	2,11	2,02	1,92	1,80
60	2,50	2,35	2,20	2,12	2,03	1,94	1,84	1,73	1,60
120	2,34	2,19	2,03	1,95	1,86	1,76	1,66	1,53	1,38
∞	2,18	2,04	1,88	1,79	1,70	1,59	1,47	1,32	1,00

B10 De toets van Grubbs

Uitschieterstoets voor enkelvoudige uitkomsten en steekproefgemiddelden.
Voorwaarde: Waarnemingen zijn normaal verdeeld.
Gegeven zijn de kritieke waarden met bijbehorende rechteroverschrijdingskansen.
Toetsingsgrootheid: $T = \dfrac{x_{(n)} - \bar{x}}{s}$ of $T = \dfrac{\bar{x} - x_1}{s}$

Kritieke waarden voor α (rechtseenzijdig)

aantal waarnemingen (n)	0,05	0,025	0,01
3	1,15	1,15	1,15
4	1,46	1,48	1,49
5	1,67	1,71	1,75
6	1,82	1,89	1,94
7	1,94	2,02	2,10
8	2,03	2,13	2,22
9	2,11	2,21	2,32
10	2,18	2,29	2,41
11	2,23	2,36	2,48
12	2,29	2,41	2,55
13	2,33	2,46	2,61
14	2,37	2,51	2,66
15	2,41	2,55	2,71
16	2,44	2,59	2,75
17	2,47	2,62	2,79
18	2,50	2,65	2,82
19	2,53	2,68	2,85
20	2,56	2,71	2,88
21	2,58	2,73	2,91
22	2,60	2,76	2,94
23	2,62	2,78	2,96
24	2,64	2,80	2,99
25	2,66	2,82	3,01
30	2,75	2,91	
35	2,82	2,98	
40	2,87	3,04	
45	2,92	3,09	
50	2,96	3,13	
60	3,03	3,20	
70	3,09	3,26	
80	3,14	3,31	
90	3,18	3,35	
100	3,21	3,38	

B11 De toets van Cochran

Uitschieterstoets voor een extreem grote variantie binnen een groep van k varianties uit (bij benadering) normale verdelingen. Elke variantie is gebaseerd op ν vrijheidsgraden.

Toetsingsgrootheid: $T = \dfrac{s^2_{\max}}{s^2_1}$

Kritieke waarden voor $\alpha = 0{,}05$ (rechtseenzijdig)

v	1	2	3	4	5	6	7	8	9	10	16	36
2	0,998	0,975	0,939	0,906	0,877	0,853	0,833	0,816	0,801	0,788	0,734	0,660
3	0,967	0,871	0,798	0,746	0,707	0,677	0,653	0,633	0,617	0,602	0,547	0,475
4	0,906	0,768	0,684	0,629	0,590	0,560	0,536	0,518	0,502	0,488	0,437	0,372
5	0,841	0,684	0,598	0,544	0,506	0,478	0,456	0,439	0,424	0,412	0,364	0,307
6	0,781	0,616	0,532	0,480	0,445	0,418	0,398	0,382	0,368	0,357	0,314	0,261
7	0,727	0,561	0,480	0,431	0,397	0,373	0,354	0,338	0,326	0,315	0,276	0,228
8	0,680	0,516	0,438	0,391	0,359	0,336	0,319	0,304	0,293	0,283	0,246	0,202
9	0,638	0,477	0,403	0,358	0,329	0,307	0,290	0,277	0,266	0,257	0,223	0,182
10	0,602	0,445	0,373	0,331	0,303	0,282	0,267	0,254	0,244	0,235	0,203	0,166
12	0,541	0,392	0,326	0,288	0,262	0,244	0,230	0,219	0,210	0,202	0,174	0,140
15	0,471	0,335	0,276	0,242	0,219	0,203	0,191	0,182	0,174	0,167	0,143	0,114
20	0,389	0,270	0,220	0,192	0,173	0,160	0,150	0,142	0,136	0,130	0,111	0,088
24	0,343	0,235	0,190	0,166	0,149	0,137	0,129	0,122	0,116	0,111	0,094	0,074
30	0,293	0,198	0,159	0,138	0,124	0,114	0,106	0,100	0,096	0,092	0,077	0,060
40	0,237	0,158	0,126	0,108	0,097	0,089	0,083	0,078	0,074	0,071	0,060	0,046
60	0,174	0,113	0,090	0,076	0,068	0,062	0,058	0,055	0,052	0,050	0,041	0,032

Kritieke waarden voor $\alpha = 0{,}01$ (rechtszijdig)

v	1	2	3	4	5	6	7	8	9	10	16	36
2	0,100	0,995	0,980	0,959	0,937	0,917	0,899	0,882	0,867	0,854	0,795	0,707
3	0,993	0,942	0,883	0,834	0,793	0,761	0,734	0,711	0,691	0,674	0,606	0,515
4	0,968	0,864	0,781	0,721	0,676	0,641	0,613	0,590	0,570	0,554	0,488	0,406
5	0,928	0,788	0,696	0,633	0,588	0,553	0,526	0,504	0,485	0,470	0,409	0,335
6	0,883	0,722	0,626	0,564	0,520	0,487	0,461	0,440	0,423	0,408	0,353	0,286
7	0,838	0,664	0,569	0,508	0,466	0,435	0,410	0,391	0,375	0,362	0,310	0,249
8	0,794	0,615	0,521	0,463	0,423	0,393	0,370	0,352	0,337	0,325	0,278	0,221
9	0,754	0,573	0,481	0,425	0,387	0,359	0,338	0,321	0,307	0,295	0,251	0,199
10	0,718	0,536	0,447	0,393	0,357	0,331	0,311	0,294	0,281	0,270	0,230	0,181
12	0,653	0,475	0,392	0,343	0,310	0,286	0,268	0,254	0,242	0,232	0,196	0,154
15	0,575	0,407	0,332	0,288	0,259	0,239	0,223	0,210	0,200	0,192	0,161	0,125
20	0,480	0,330	0,265	0,229	0,205	0,188	0,175	0,165	0,157	0,150	0,125	0,096
24	0,425	0,287	0,230	0,197	0,176	0,161	0,150	0,141	0,134	0,128	0,106	0,081
30	0,363	0,241	0,191	0,164	0,145	0,133	0,123	0,116	0,110	0,105	0,087	0,066
40	0,294	0,192	0,151	0,128	0,114	0,103	0,096	0,090	0,085	0,082	0,067	0,050
60	0,215	0,137	0,107	0,090	0,078	0,072	0,067	0,062	0,059	0,057	0,046	0,034

Bij een gering verschil in aantal vrijheidsgraden neemt men het gemiddelde van deze aantallen.

B12 Constanten voor berekening van lijnen op controlekaarten

	$\bar{x}$-kaart μ en σ niet gegeven		R-kaart σ niet gegeven		R-kaart σ gegeven			s-kaart σ gegeven			s-kaart σ niet gegeven	
bovengrens	$\bar{x} + A_2\bar{R}$	$\bar{x} + A_3\bar{S}$	$D_4\bar{R}$		$D_2\sigma$			$B_2\sigma$			$B_4\bar{s}$	
controlelijn	$\bar{x}$	$\bar{x}$	$\bar{R}$		$d_2\sigma$			$c_4\sigma$			$\bar{s}$	
ondergrens	$\bar{x} - A_2\bar{R}$	$\bar{x} - A_3\bar{S}$	$D_3\bar{R}$		$D_1\sigma$			$B_1\sigma$			$B_3\bar{s}$	
n	A_2	A_3	D_3	D_4	d_2	D_1	D_2	c_4	B_1	B_2	B_3	B_4
2	1,880	2,659	0	3,267	1,128	0	3,686	0,7979	0	1,843	0	3,267
3	1,023	1,954	0	2,575	1,693	0	4,358	0,8862	0	1,858	0	2,568
4	0,729	1,628	0	2,282	2,059	0	4,698	0,9213	0	1,808	0	2,266
5	0,577	1,427	0	2,115	2,326	0	4,918	0,9400	0	1,756	0	2,089
6	0,483	1,287	0	2,004	2,534	0	5,078	0,9515	0,026	1,711	0,030	1,970
7	0,419	1,182	0,076	1,924	2,704	0,205	5,203	0,9594	0,105	1,672	0,118	1,882
8	0,373	1,099	0,136	1,864	2,847	0,387	5,307	0,9650	0,167	1,638	0,185	1,815
9	0,337	1,032	0,184	1,816	2,970	0,546	5,394	0,9693	0,219	1,609	0,239	1,761
10	0,308	0,975	0,223	1,777	3,078	0,687	5,469	0,9727	0,262	1,584	0,284	1,716
11	0,285	0,927	0,256	1,744	3,173	0,812	5,534	0,9754	0,299	1,561	0,321	1,679
12	0,266	0,886	0,284	1,716	3,258	0,924	5,592	0,9776	0,331	1,541	0,354	1,646
13	0,249	0,850	0,308	1,692	3,336	1,026	5,646	0,9794	0,359	1,523	0,382	1,618
14	0,235	0,817	0,329	1,671	3,407	1,121	5,693	0,9810	0,384	1,507	0,406	1,594
15	0,223	0,789	0,348	1,652	3,472	1,207	5,737	0,9823	0,406	1,492	0,428	1,572

B12 (vervolg)

	$\bar{x}$-kaart μ en σ niet gegeven		R-kaart σ niet gegeven		R-kaart σ gegeven			s-kaart σ gegeven			s-kaart σ niet gegeven	
bovengrens	$\bar{x} + A_2\bar{R}$	$\bar{x} + A_3\bar{S}$	$D_4\bar{R}$		$D_2\sigma$			$B_2\sigma$			$B_4\bar{s}$	
controlelijn	$\bar{x}$	$\bar{x}$	$\bar{R}$		$d_2\sigma$			$c_4\sigma$			$\bar{s}$	
ondergrens	$\bar{x} - A_2\bar{R}$	$\bar{x} - A_3\bar{S}$	$D_3\bar{R}$		$D_1\sigma$			$B_1\sigma$			$B_3\bar{s}$	
n	A_2	A_3	D_3	D_4	d_2	D_1	D_2	c_4	B_1	B_2	B_3	B_4
16	0,212	0,763	0,364	1,636	3,532	1,285	5,779	0,9835	0,427	1,478	0,448	1,552
17	0,203	0,739	0,379	1,621	3,588	1,359	5,817	0,9845	0,445	1,465	0,466	1,534
18	0,194	0,718	0,392	1,608	3,640	1,426	5,854	0,9854	0,461	1,454	0,482	1,518
19	0,187	0,698	0,404	1,596	3,689	1,490	5,888	0,9862	0,477	1,443	0,497	1,503
20	0,180	0,680	0,414	1,586	3,735	1,548	5,922	0,9869	0,491	1,433	0,510	1,490
21	0,173	0,663	0,425	1,575	3,778	1,606	5,950	0,9876	0,504	1,424	0,523	1,477
22	0,167	0,647	0,434	1,566	3,819	1,659	5,979	0,9882	0,516	1,415	0,534	1,466
23	0,162	0,633	0,443	1,557	3,858	1,710	6,006	0,9887	0,527	1,407	0,545	1,455
24	0,157	0,619	0,452	1,548	3,895	1,759	6,031	0,9892	0,538	1,399	0,555	1,445
25	0,135	0,606	0,459	1,541	3,931	1,804	6,058	0,9896	0,548	1,392	0,565	1,435

Bijlage C Antwoorden

Hoofdstuk 3

1. **a.** $x_5 = 115 - 96 = 19$ **b.** $s^2 = \frac{256}{4} = 64$

2. **a.** Me $= 52{,}35$; $\bar{x} = 52{,}89$ **b.** $R = 5$; $s = 1{,}65$
c. $c = 3{,}1\%$

3. **a.** $\bar{x} = 35$; $s = 9{,}27$ **b.** $c = 26\%$

4. Als Y is °F$\Longrightarrow \bar{y} = 129{,}7$; $s_Y = 4{,}23$ en $c_y = 3{,}3\%$

5. **a.** $\mu = 22{,}5$ en $\sigma = 2{,}4$ **b.** $\mu = \frac{25}{3}$ en $\sigma = 0{,}8$
c. $\mu = 27{,}5$ en $\sigma = 2{,}4$ **d.** $\mu = 75$ en $\sigma = 7{,}2$
e. $\mu = 7{,}5$ en $\sigma = 0{,}8$ **f.** $\mu = \frac{35}{6}$ en $\sigma = 0{,}8$
g. $\mu = 82{,}5$ en $\sigma = 7{,}2$ **h.** $\mu = 77{,}5$ en $\sigma = 7{,}2$

6.

Klassen	Klassenmidden	Frequentie
6,0 - < 9,0	7,5	3
9,0 - <12,0	10,5	8
12,0 - <15,0	13,5	9
15,0 - <18,0	16,5	12
18,0 - <21,0	19,5	20
21,0 - <24,0	22,5	12
24,0 - <27,0	25,5	11
27,0 - <30,0	28,5	4
30,0 - <33,0	31,5	1

$\bar{x} = 18{,}9$ en $s = 5{,}6$

7. **a.** Mo $= 47$; $\bar{x} = 56{,}9$; Me $= 54{,}8$ **b.** $R = 71$;

c.

Klassen	Klassenmidden	Frequentie
30 -< 40	35	2
40 - < 50	45	13
50 - < 60	55	17
60 - < 70	65	10
70 - < 80	75	5
80 - < 90	85	2
90 - <100	95	-
100 - <110	105	1

d. Mo $= 55$; $\bar{x} = 55{,}6$; Me $= 55{,}9$

e. De sterk uitschietende waarde 105 verstoort het "werkelijke" beeld

8. a.

Klassen	frequentie
0,60 - <0,63	2
0,63 - <0,66	4
0,66 - <0,69	10
0,69 - <0,72	22
0,72 - <0,75	7
0,75 - <0,78	3
0,78 - <0,81	2

b. $\bar{x} = 0{,}702$ en $s = 0{,}039$

9. a.

Klassen	freq.	rel. freq.	rel. cum. freq.
0 - <4	3	0,06	0,06
4 - <8	11	0,22	0,28
8 - <12	14	0,28	0,56
12-<16	9	0,18	0,74
16-<20	6	0,12	0,86
20-<24	2	0,04	0,90
24-<28	2	0,04	0,94
28-<32	3	0,06	1,00

c. $\bar{x} = 12{,}64$ en $s = 7{,}2$

d. Me = 11,1 en Mo = 10

Hoofdstuk 4

1. **a.** $P(M) = \frac{4}{10}$; $P(50 \text{ of ouder}) = \frac{3}{10}$; $P(M \mid \text{wel kinderen}) = \frac{3}{6}$; $P(\text{wel kinderen} \mid M) = \frac{3}{4}$; $P(V \mid \text{onder de } 50) = \frac{5}{7}$; $P(\text{onder de } 50 \mid V) = 1$
   b. $P(M \cap \text{geen kinderen}) = \frac{1}{10}$
   c. $P(\text{boven } 50 \cap \text{kinderen} \mid M) = \frac{3}{4}$
   d. $P(V \cup \text{boven } 50) = \frac{9}{10}$
   e. $\frac{2}{15}$
   f. $\frac{8}{15}$
   g. $\frac{1}{15}$

2. **a.** $\frac{53}{54}$
   b. $\frac{5}{72}$

3. **a.** $\frac{5}{12}$
   b. $\frac{1}{4}$
   c. $\frac{1}{12}$
   d. $\frac{7}{12}$

4. **a.** $\frac{1}{17}$
   b. $\frac{13}{102}$
   c. $\frac{2}{17}$

5. $0{,}8 \times 0{,}7 + 0{,}5 \times 0{,}05 + 0{,}1 \times 0{,}25 = 0{,}61$

6. **a.** $0{,}2 \times 0{,}7 + 0{,}8 \times 0{,}9 = 0{,}86$
   b. $\frac{0{,}72}{0{,}86} = 0{,}8372$

7. $\frac{1000}{0{,}12} \approx 8334$

8. **a.** $0{,}2 \times 0{,}03 + 0{,}3 \times 0{,}05 + 0{,}5 \times 0{,}1 = 0{,}071$
b. $\frac{0{,}05}{0{,}071} = \frac{50}{71}$

9. **a.** $0{,}1 \times 0{,}95 = 0{,}095$
b. $0{,}095 + 0{,}08 \times 0{,}9 = 0{,}177$
c. $\frac{0{,}9 \times 0{,}08}{0{,}177} = 0{,}4068$

10. **a.** $\frac{6}{11}\frac{7}{10} + \frac{3}{11}\frac{5}{10} + \frac{2}{11}\frac{9}{10} = \frac{15}{22}$
b. $\frac{\frac{42}{110}}{\frac{15}{22}} = \frac{15}{22}$

11. **a.** $\frac{11}{15}$
b. $\frac{3}{5}$
c. $\frac{2}{5}$
d. nee want $\frac{11}{15} \times \frac{3}{5} \neq \frac{2}{5}$
e. $\frac{6}{11}$

12. **a.** $8 \times 7 = 56$
b. $2 \times (6 \times 2) = 24$
c. 12

13. **a.** $\frac{7!}{2!2!} = 1260$
b. $\frac{10!}{5!2!} = 15120$

14. $\binom{4}{2}^2 \times 4! = 864$

15. **a.** 1 op $\binom{45}{6}$ = 1 op 8145060
b. $\frac{\binom{6}{5}\binom{39}{1}}{\binom{45}{6}}$ = 234 op 8145060

16. $\binom{10}{3}\left(\frac{1}{3}\right)^3\left(\frac{2}{3}\right)^7 = 0{,}2601$

17. $1 - \left\{1 \times \frac{364}{365} \times \frac{363}{365} \times \frac{362}{365} \times \ldots\ldots \times \frac{346}{365}\right\} = 0{,}41$

18. **a.** $(0{,}95)^{10} = 0{,}5987$
b. $0{,}5987 + 10(0{,}05)(0{,}95)^9 = 0{,}9138$
c. $1 - \left\{0{,}9138 + \binom{10}{2}(0{,}05)^2(0{,}95)^8\right\} = 0{,}0116$

19. **a.** $\frac{14}{20}\frac{13}{19}\frac{12}{18}\frac{11}{17}\frac{10}{16} = 0{,}1291$
b. $0{,}1291 + 5\frac{6}{20}\frac{14}{19}\frac{13}{18}\frac{12}{17}\frac{11}{16} = 0{,}5165$
c. $1 - (0{,}5165 + \binom{6}{2}\frac{6}{20}\frac{5}{19}\frac{14}{18}\frac{13}{17}\frac{12}{16} = 0{,}1313$

20. **a.** $(0{,}3)^4 = 0{,}0081$
b. $4!(0{,}3)(0{,}2)(0{,}1)(0{,}4) = 0{,}0576$
c. $\binom{4}{2}(0{,}3)^2(0{,}7)^2 = 0{,}2646$

Hoofdstuk 5

1. $P(K \geq 1) = 1 - P(K = 0) = 1 - 0{,}3164 = 0{,}6836$

2. $P(K = 1) = 0{,}3932$

3. $P(K \leq 1) = 0{,}4450$

4. $P(K > 2) = 1 - P(K \leq 2) = 1 - 0{,}6769 = 0{,}3231$

5. **a.** $P(K \geq 1|n = 15 \text{ en } p = 0{,}05) = 1 - P(K = 0) = 1 - 0{,}4633 = 0{,}5367$
b. $P(K = 0)|n = 15 \text{ en } p = 0{,}10) = 0{,}2059$

6. $P = 0{,}5972$

7. **a.** $P(K = 2) = 0{,}2293$ **b.** $P(K \leq 5) = 0{,}9327$
c. $P(K \leq 2) = 0{,}4049$

8. $P(K = 2) = 0{,}0164$

9. **a.** $P(K \leq 2) = 0{,}0620$
b. $P(K \geq 2) = 1 - P(K \leq 1) = 1 - 0{,}5578 = 0{,}4422$
c. $P(5 \leq K \leq 8) = P(K \leq 8) - P(K \leq 4) = 0{,}4557 - 0{,}0550 = 0{,}4007$

10. $P(K \geq 10) = 1 - P(K \leq 9) = 1 - 0{,}7764 = 0{,}2236$

11. **a.** $P(K = 0|\lambda = 0{,}5) = 0{,}6065$ **b.** $P(K = 0|\lambda = 1{,}5) = 0{,}2232$
c. $P(K \geq 2|\lambda = 1{,}5) = 1 - P(K \leq 1) = 1 - 0{,}5578 = 0{,}4422$

12. $m = 8$

13. **a.** $P(K = 0|$ hypergeometrische verdeling$) = 0{,}40$ **b.** $P(K = 1) = 0{,}45$
c. $P(K \geq 2) = 1 - P(K \leq 1) = 1 - \{0{,}40 + 0{,}45\} = 0{,}15$

14. **a.** $P(K = 0|$ hypergeometrische verdeling$) = 0{,}18$ **b.** $P(K = 5) = 0{,}073$

Hoofdstuk 6

1. **a.** $P(X < 53{,}2) = P(U < -0{,}59) = 0{,}2776$ **b.** $P(X > 70) = P(U > 0{,}25) = 0{,}4013$
c. $P(83{,}2 < X < 95{,}7) = P(X > 83{,}2) - P(X > 95{,}7) = 0{,}1814 - 0{,}0618 = 0{,}1196$
d. $P = 0{,}1506$ **e.** $P = 0{,}0952$
f. $P = 0{,}7566$

2. $P = 0{,}0475$

3. $\mu = 255{,}825$

4. $\sigma = 5{,}0$

5. 438,85 euro

6. 17,50 mm

7. $P(K > 31|\text{Bin}(n = 225 : p = 0{,}1)) = P(X > 31{,}5|N(\mu = 22{,}5; \sigma^2 = 20{,}25) = P(U > 2) = 0{,}0228$

8. Kans op overstromen: $P(X > 108) = P(U > 1{,}96) = 0{,}025$
Bij 100 maal vullen: $P(K \geq 1) = 1 - P(K = 0|\ \text{Bin}(n = 100; p = 0{,}025)) = 1 - 0{,}0795 = 0{,}9205$

9. $P(K \leq 50|\ \text{Bin}(n = 100; p = 0{,}60)) = P(X < 50{,}5|N(\mu = 60; \sigma^2 = 24) = P(U < -1{,}94) = 0{,}0262$

10. $e^{-\frac{1}{3}}$=0,7165

11. $\mu_t = 2$ min/gesprek $\Rightarrow \lambda = \frac{1}{2}$ gesprek/min $\Rightarrow f(t) = \frac{1}{2}e^{-\frac{t}{2}}$
   a. $P(T > 2) = \int_2^{\infty} \frac{1}{2}e^{-\frac{t}{2}}dt = [-e^{-\frac{t}{2}}]_2^{\infty} = e^{-1} = 0{,}3679$
   b. $E(K) = 0{,}1 \cdot P(T \leq 2) + 0{,}2 \cdot P(T > 2) = 0{,}1 \cdot \{1 - P(T > 2)\} + 0{,}2 \cdot P(T > 2)$
   $E(K) = 0{,}1 \cdot \{1 - 0{,}3679\} + 0{,}2 \cdot 0{,}3679 = 0{,}137$ euro

Hoofdstuk 7

1. 0,0287

2. 0,0548

3. a. 0,0158 b. 0,1814
   c. 0,0272

4. a. $\mu_K = 106$ kg en $\sigma_K = 30$ kg b. $\mu_B = 220{,}5$ kg

5. a. $\mu_K = 3{,}5$ en $\sigma_K = 1{,}701$ b. $\mu_L = 7$ en $\sigma_L = 2{,}415$
   c. $\mu_M = 10{,}5$ en $\sigma_M = 2{,}946$

6. a. $\mu = 30$ dagen en $\sigma = 3$ dagen b. 44,12 dagen en 55,88 dagen
   c. 38,63 dagen d. 61,37 dagen

7. 2,28%

8. a. $\mu = 7000$ gram en $\sigma = 65$ gram b. 0,2758
   c. 5 g

9. 466 doosjes

10. **a.** $\sigma = 1$ mm
b. 12,92%
c. 20%
d. $\mu_P = 3{,}4$ mm en $\sigma_P = 0{,}4$ mm
e. 5%

11. **a.** 26,6%
b. 3784 uur
c. 10,56%
d. 3392 uur
e. 16 lampjes

12. 0,1003

13. 0,8413

14. $\mu = 25$ mm en $\sigma = 2$ mm

15. **a.** 0,2119
b. 0,9876

16. **a.** 0,2743
b. 100% kans dat $M_A < M_B$

Hoofdstuk 8

1. **a.** 0,970
b. 0,030
c. 1,812; −2,228; −1,372; 2,764

2. **a.** 0,05
b. 0,05
c. 0,4101
d. 28,87; 7,02
e. 69,6; 62,8

3. **a.** [98,90; 109,30]
b. [99,71; 108,49]

4. [89,38; 90,62]

5. [8,47; 10,50]

6. [2,13; 10,83]

7. **a.** [0,644; 0,756]
b. 385
c. [0,643; 0,757]

8. [30,33%; 36,47%]

9. **a.** [25,35; 87,29]
b. ja

10. [65,97%; 69,97%]; [14,49%; 17,64%]; [5,06%; 7,10%]; [8,97%; 11,36%]

Hoofdstuk 9

1. Hypothesen: H_0: $\mu = 112$ en H_1: $\mu < 112$ ($\alpha = 0{,}05$ eenzijdig)
u-toets (σ bekend) $\Rightarrow P(U < -2{,}85) = 0{,}0022$. H_0 verwerpen; de broeken zijn te kort.

2. Hypothesen: H_0: $\mu = 7{,}1$ en H_1: $\mu > 7{,}1$ ($\alpha = 0{,}05$ eenzijdig)
t-toets (σ onbekend) $\Rightarrow P(T < -2{,}74)$ bij $\nu = 9$. De overschrijdingskans: $0{,}01 < P < 0{,}025$. H_0 verwerpen.

3. t-toets voor twee onafhankelijke steekproeven.
Hypothesen: H_0: $\mu_a = \mu_b$ en H_1: $\mu_a \neq \mu_b$ ($\alpha = 0{,}05$ tweezijdig)
Eerst toetsen of s_1^2en s_2^2 schattingen zijn van dezelfde σ^2 m.b.v. de F-toets: $F[4, 5] = \frac{6{,}16}{5{,}30} = 1{,}16 \Rightarrow$ $P(F > 1{,}16) > 0{,}05$. De beide steekproefvarianties zijn schattingen van dezelfde σ^2.
s^2 berekenen als een gewogen gemiddelde van s_1^2en s_2^2: $\frac{4\times 5{,}30+5\times 6{,}16}{4+5} = 5{,}76 \Rightarrow s = \sqrt{5{,}76} = 2{,}4$, met $\nu = 9$ vrijheidsgraden.
T heeft waarde $\frac{61{,}82-58{,}60}{2{,}4\sqrt{\frac{1}{5}+\frac{1}{6}}} = 2{,}22$
$\Rightarrow 0{,}025 < P(T < 2{,}22) < 0{,}05$ (eenzijdig)
$\Rightarrow 0{,}05 < P(T < 2{,}22) < 0{,}10$ (tweezijdig)
H_0 niet verwerpen. Geen verschil aangetoond tussen beide apparaten.

4. t-toets voor twee gepaarde steekproeven.
Hypothesen: H_0: $\mu_V = 0$ en H_1: $\mu_V \neq 0$ (met $V = B - A$) ($\alpha = 0{,}05$ tweezijdig)
$\bar{V} = 1{,}89$, $s_V = 7{,}56$ en $\nu = 8 - 1 = 7$.
T heeft waarde $\frac{1{,}89}{\frac{2{,}76}{\sqrt{8}}} = 1{,}94$
$\Rightarrow 0{,}025 < P(T > 1{,}94) < 0{,}05$ (eenzijdig)
$\Rightarrow 0{,}05 < P(T > 1{,}94) < 0{,}10$ (tweezijdig)
H_0 niet verwerpen. Geen verschil aangetoond tussen beide apparaten.

5. F-Toets voor twee varianties. Hypothesen: H_0: $\sigma_1^2 = \sigma_2^2$ en H_1: $\sigma_1^2 \neq \sigma_2^2$.
F heeft waarde $\frac{2{,}87^2}{1{,}44^2} = 3{,}97$, met $\nu_1 = 9$ en $\nu_2 = 9$
$\Rightarrow P(F > 3{,}97) > 0{,}05$ (eenzijdig)
$\Rightarrow P(F > 3{,}97) > 0{,}10$ (tweezijdig)
H_0 niet verwerpen. Geen verschil aangetoond tussen beide analisten.

6. Uitschieters: 1,45 en/of 1,10? $\bar{X} = 1{,}259$ en $s = 0{,}089$.
$\Rightarrow 1{,}45$: T heeft waarde $\frac{1{,}45-1{,}259}{0{,}089} = 2{,}15$.
$\Rightarrow 0{,}025 < P(T > 2{,}15) < 0{,}05$. 1,45 is uitschieter, maar geen storende uitschieter.
$\Rightarrow 1{,}10$: T heeft waarde $\frac{1{,}259-1{,}10}{0{,}089} = 1{,}79$.
$\Rightarrow P(T > 2{,}15) > 0{,}05$. 1,10 is geen uitschieter.

7. $s_{\max}^2 = 79{,}2$ en $\sum_{i=1}^{k} s_i^2 = 218{,}4$
$\Rightarrow T = \frac{79{,}2}{218{,}4} = 0{,}363$
$\Rightarrow P(T > 0{,}363) > 0{,}05$. 79,2 is geen uitschieter.

8. $s_A^2 = 0{,}043$, $s_B^2 = 0{,}312$, $s_C^2 = 0{,}112$, $s_D^2 = 0{,}475$, $s_E^2 = 2{,}572$.
$\sum_{i=1}^{k} s_i^2 = 3{,}514 \Rightarrow T = \frac{2{,}572}{3{,}514} = 0{,}732$

$\Rightarrow P(T > 0{,}732) < 0{,}01$. Steekproef E heeft een te grote variantie t.o.v. de andere steekproeven. Dit ontstaat door de waarde 2,9 in steekproef E.

9. **a.** T is het steekproefgemiddelde van vier waarnemingen.
Onder H_0 volgt T een normale verdeling $\mu = 3{,}5$ en $\sigma = \frac{2}{\sqrt{4}} = 1$.

b. $P(T > 5{,}46) = P(U > \frac{5{,}46-3{,}5}{1}) = P(U > 1{,}96) = 0{,}025$

c. Onder H_1 volgt T een normale verdeling $\mu = 6{,}75$ en $\sigma = \frac{2}{\sqrt{4}} = 1$.

d. $P(T > 5{,}46|\mu = 6{,}75) = P(U > -1{,}29) = 1 - P(U < -1{,}29) = 1 - 0{,}0985 = 0{,}9015$

e. De waarde van T is 5,46 is de kritieke waarde.

f. Het gebied rechts van 5,46 is het kritieke gebied

g. De onbetrouwbaarheid van de toets α is in vraag b berekend: $\alpha = 0{,}025$

h. Het onderscheidingsvermogen $1 - \beta$, is de kans dat $T \geq 5{,}46$ indien H_1 waar is.
Dit is in vraag d berekend: $1 - \beta = 0{,}9015$.

10. Chi-kwadraattoets. Toetsingsvariabele C met waarde 24,54 met $\nu = 1 \times 1 = 1$.
$P(C > 24{,}54|\ \chi^2[1]) < 0.005$
H_0 verwerpen. Er bestaat een effect door de reclamecampagne.

11. Chi-kwadraattoets. Toetsingsvariabele C met waarde1,029, met $\nu = 1 \times 2 = 2$.
$P(C > 1{,}029|\chi^2[2]) > 0{,}10$.
H_0 niet verwerpen. Er bestaat geen verschil in besmetting bij de methoden A, B en C.

12. Hypothesen: H_0: $\mu = 8000$ en $H_1 : \mu > 8000$. t-toets voor één gemiddelde.
Toetsingsvariabele T met waarde $\frac{8300-8000}{1000/\sqrt{20}} = 1{,}342$, met $\nu = 19$.
$P(T > 1{,}342) > 0{,}10$. H_0 niet verwerpen. De levensduur is niet langer dan 8000 uur.

13. Hypothesen: H_0: Er is geen verband tussen de afdeling en de werkomstandigheden, H_1: Er is wel een samenhang tussen de afdeling en de werkomstandigheden. Chi-kwadraattoets. Toetsingsvariabele C met waarde 24,32, met $\nu = 2 \times 2 = 4$.
$P(T > 24{,}32 \mid \chi^2[4]) < 0{,}01$.
H_0 verwerpen, er bestaat een verband tussen de afdeling en de werkomstandigheden.

14. Hypothesen: H_0: $\mu = 500$ en H_1: $\mu < 500$. u-toets voor één gemiddelde.
Toetsingsvariabele U met waarde $\frac{485-500}{\frac{28}{\sqrt{16}}} = 2{,}143$.
$P(U > 2{,}14) > 0{,}0162$. H_0 verwerpen. Proces is verschoven, machine bijstellen.

15. H_0: $\mu_A = \mu_B$ en H_1: $\mu_A \neq \mu_B$. u-toets want de σ's zijn bekend.
Toetsingsvariabele U met waarde $\frac{60000-59000}{\sqrt{\frac{1200^2}{12} + \frac{1600^2}{15}}} = 1{,}84$.
$P(U > 1{,}84) > 0{,}0329$ (Eenzijdig)
$P(U > 1{,}84) > 0{,}0658$ (Tweezijdig)
H_0 niet verwerpen. Er is (net) geen verschil tussen beide methoden aangetoond.

16. H_0: $\sigma^2_{na} = \sigma^2_{voor}$ en H_1: $\sigma^2_{na} > \sigma^2_{voor}$. F-toets voor 2 varianties.
Toetsingsvariabele $F[15, \infty]$ met waarde $\frac{24^2}{20^2} = 1{,}44$.
$P(F > 1{,}84|\ \nu_1 = 15);\ \nu_2 = \infty) > 0{,}05$ (Eenzijdig)
H_0 niet verwerpen. De spreiding is niet groter geworden.

17. $H_0 : p_A = p_B = 0{,}5$ en $H_1 : p_A > 0{,}5$. K =aantal personen met voorkeur voor soep A.
Onder H_0: $K \backsim \text{Bin}(n = 100 \text{ en } p = 0{,}5) \Rightarrow N(\mu = 50; \sigma^2 = 25)$.
$P(K \geq 63 \mid \text{Bin.}) = P(X > 62{,}5 \mid \text{Norm.}) \Rightarrow P(U > \frac{62{,}5-50}{5}) = P(U > 2{,}50) = 0{,}0062$.
H_0 verwerpen. De voorkeur voor soep A is groter dan voor soep B.

Hoofdstuk 10

1. $\hat{y} = 3\hat{x}$; $\sum r_i = 0$; $\sum(r_i)^2 = 24$

2. $\hat{x} = 2y + 3$; $\sum s_i = 0$; $\sum(s_i)^2 = 0$

3. **a.** $\hat{y} = 1{,}76x + 14{,}53$ **b.** $\hat{x} = 0{,}00375y + 0{,}49$
c. $\rho = 0{,}08$

4. **a.** $\hat{z} = 1{,}86d - 1{,}78$; $\hat{d} = 0{,}53z + 1{,}06$ **b.** 0,996
c. ja $(21{,}4 >> 1{,}86 \times 11{,}7 - 1{,}78 + 3s_r = 19{,}98)$

5. **a.** $\hat{y} = 0{,}00753x + 1{,}2533$; $\hat{x} = 71{,}25y + 97{,}5$ **b.** $r = 0{,}732$

6. $Z = -86{,}9727 - 12{,}3757X + 146{,}6868Y$

7. $a = 30{,}36$ en $b = 1{,}445$

8. **b.** $a = 0{,}116$ en $b = 0{,}245$ **c.** $a = 0{,}88$ en $b = 1{,}36$

9. 0,318

10. **a.** 9,5% **b.** 4,56%
c. −0,375 **d.** 0,08%
e. 0,60 **f.** 0,80

Hoofdstuk 11

1. $89 \pm 3\frac{1{,}5}{\sqrt{3}} = 89 \pm 2{,}6$

2. $\bar{\bar{x}} = 319{,}6$ en $\bar{s} = 7{,}82$
$\bar{x}$-kaart: Bovengrens: 332,4; norm: 319,6; ondergrens: 306,9
s-kaart: Bovengrens: 17,7; norm: 7,8; ondergrens: 0.

3. $\bar{x} = 322{,}3$ en $\bar{R} = 9{,}7$
$\bar{x}$-kaart: Bovengrens: 351,4; norm: 322,3; ondergrens: 293,2
R-kaart: Bovengrens: 35,8; norm: 10,9; ondergrens: 0.

4. $\bar{\bar{x}} = 50{,}06$ en $\bar{s} = 0{,}24$
$\bar{x}$-kaart: Bovengrens: 50,53; norm: 50,06; ondergrens: 49,58
s-kaart: Bovengrens: 0,63; norm: 0,24; ondergrens: 0.

8. $C_p = 1{,}13$ en $C_{pk} = 1{,}13$. Zowel niveau als spreiding ok.

9. Capability van het proces: $\hat{\mu} \pm A_2\bar{R} = 255 \pm 0{,}729 \times 11 = 255 \pm 8$
$C_{pk} = \frac{USL-\mu}{3\sigma} = 1{,}2 \Rightarrow USL - 255 = 3 \times 4 \Rightarrow USL = 267$ en $LSL = 243$

10. **a.** $C_p = 0{,}44$ en $C_{pk} = 0{,}38$

b. $P(X > 35{,}6|\ \mu = 32{,}5; \sigma = 2{,}7) = P(U > 1{,}15) = 20{,}59\%$

$P(X < 28{,}5|\mu = 32{,}5; \sigma = 2{,}7) = P(U < -1{,}48) = 13{,}34\%$

Totaal 33,93% van de productie buiten de specificatiegrenzen.

Register

Statistiek is in onze moderne samenleving niet meer weg te denken. In de handel en industrie, bij de overheid, in het bank- en verzekeringswezen, en vooral ook in de wetenschap speelt het een belangrijke rol. Om meetresultaten op de juiste wijze te kunnen interpreteren en te analyseren, moeten vrijwel alle studenten in het hoger onderwijs uitgerust zijn met een behoorlijke hoeveelheid basiskennis van de statistiek.

Dit boek biedt een overzicht van alle basisbeginselen van de statistiek en de voornaamste toepassingen daarvan, zowel op het gebied van de beschrijvende als van de toegepaste statistiek. Ingespeeld wordt op de veranderingen in het wiskunde- en statistiekonderwijs in het HBO, met het accent op toepassingen in de praktijk. De auteurs besteden ruime aandacht aan statistische toepassingen in de kwaliteitskunde en procesbeheersing. Aan de in de praktijk meest gebruikte toetsen is een apart hoofdstuk gewijd.

Het boek leent zich uitstekend voor gebruik in extensiverende onderwijsvormen (zelfstudie, college-instructievorm), onder andere door toevoeging van motiverende opgaven tussen de teksten en voorbeelden van toepassingen met Excel. Aan het eind van elk hoofdstuk is een aantal vraagstukken opgenomen.

ISBN 90-5574-239-2



HB Hoger onderwijs Beroepspraktijk